Introductory statistics students' conceptual understanding of study design and conclusions

A Dissertation

SUBMITTED TO THE FACULTY OF

UNIVERSITY OF MINNESOTA

BY

Elizabeth Brondos Fry

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Robert delMas, Adviser

Andrew Zieffler, Co-Adviser

December 2017

Acknowledgements

The path to completing my PhD dissertation has been a long and windy road, and I am grateful for so many individuals who have helped me along the way. First, I would like to thank Dr. Bob delMas, my primary advisor, for all of the guidance, feedback, and very careful editing throughout the process of completing my dissertation work. I am also grateful to have an excellent co-advisor, Dr. Andy Zieffler, whose help and feedback have been instrumental, especially in the development of the assessments and activities. I also want to thank Dr. Joan Garfield for starting the statistics education graduate program and for all of the advice and guidance throughout my time here. Bob, Andy, and Joan, you have helped me grow immensely as a teacher and researcher, and I hope to go on to inspire others as you have inspired me. Thank you to Dr. Lesa Clarkson for also serving on my committee, and for providing a valuable perspective from the field of math education. I also want to thank other faculty members who assisted in my research: Dr. Roxy Peck, Dr. Rob Gould, Dr. John Holcomb, and Dr. Michael Rodriguez.

I am extremely grateful for my husband Matt for putting up with me as a perpetual graduate student all of these years. I truly could not have done this without your love, patience, support, encouragement, and help. Thanks to you and Val for making sure I take time to relax. ☺ I am grateful to my parents, David and Alicia, for promoting in me the love of learning from a very young age, and to my sister Monica for your encouragement and for helping convince me to move to Minnesota. I am also thankful for the support of Matt's family. I have greatly appreciated all of the scholarly advice I have received from my father and father-in-law, both PhDs in the family.

I am grateful for the friendship and support I have received from my stat ed family, especially those whose help was indispensable for this project. Thank you, Ethan and Jonathan, for taking lots of time out of your busy schedules to help me observe class sessions. Thank you to Anelise, Nicola, and Mike for implementing the study design unit, for being patient with me throughout the development of the materials, and for all of your thoughtful feedback and tough questions. I could not have asked for a better team of instructors to implement the study design unit. Laura L, Laura Z, and Michelle, thank you also for your friendship and help, and for designing an excellent EPSY 5261 curriculum which helped me to do my job and still have time to make progress on my dissertation. Thank you also to Kory, Mario, Martin, Yadira, Astrid, and all of my other QME/PsyF friends who have supported me along the way.

Thank you to all of my friends outside of school, both near and far, for your friendship throughout the years and for staying in touch even when life is busy. I am also grateful for our friends at University Lutheran Church of Hope, who have embraced Matt and me into the church family and made us feel so welcome. And finally, I am eternally grateful to God for blessing me with talents and passions and for giving me the strength to persist.

Abstract

Recommended learning goals for students in introductory statistics courses include the ability to recognize and explain the key role of randomness in designing studies and in drawing conclusions from those studies involving generalizations to a population or causal claims (GAISE College Report ASA Revision Committee, 2016). The purpose of this study was to explore introductory statistics students' understanding of the distinct roles that random sampling and random assignment play in study design and the conclusions that can be made from each. A study design unit lasting two and a half weeks was designed and implemented in four sections of an undergraduate introductory statistics course based on modeling and simulation. The research question that this study attempted to answer is: *How does introductory statistics students' conceptual understanding of study design and conclusions (in particular, unbiased estimation and establishing causation) change after participating in a learning intervention designed to promote conceptual change in these areas?* In order to answer this research question, a forced-choice assessment called the Inferences from Design Assessment (IDEA) was developed as a pretest and posttest, along with two open-ended assignments, a group quiz and a lab assignment. Quantitative analysis of IDEA results and qualitative analysis of the group quiz and lab assignment revealed that overall, students' mastery of study design concepts significantly increased after the unit, and the great majority of students successfully made the appropriate connections between random sampling and generalization, and between random assignment and causal claims. However, a small, but noticeable portion of students continued to demonstrate misunderstandings, such as confusion between random sampling and random assignment.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statistical inference, an important component of introductory statistics courses, includes going beyond the data at hand to make a wider conclusion, which involves consideration of study design. The Guidelines for Assessment and Instruction in Statistics Education (GAISE, 2016) recommend that introductory statistics courses produce statistically educated students, who can develop statistical literacy and are able to think statistically. The GAISE guidelines outline major learning goals for students, such as being able to recognize and explain the role of randomness in study design and conclusions (GAISE, 2016, p. 10). Statistical reasoning about data includes understanding why (1) random sampling allows the results of statistical studies to be extended to the population from which the sample was generated, and (2) random assignment allows cause-and-effect conclusions to be made from comparative experiments (Garfield & Ben-Zvi, 2008, p.129-132). In order to be an educated citizen and be able to think critically about research studies, students must understand (1) common sources of bias in studies, including the lack of a representative sample, and (2) when cause and effect conclusions can be made, depending on whether a study is observational or experimental (Utts, 2003).

Thus, understanding the role of random sampling and random assignment in making inferences from statistical studies is a desired learning outcome for students of introductory statistics. However, achieving this desired learning outcome can be difficult for students of introductory statistics. Some problems have been documented in the literature about students' understanding of these ideas. Students may also enter a course with misconceptions about sampling and experimental design that tend to be difficult to

overcome (Sawilowsky, 2004; Wagler & Wagler, 2013). Another problem is that after learning about study design, students may fail to distinguish between the role of random sampling in generalization to a population, and the role of random assignment in enabling cause-and-effect conclusions to be made (Derry, Levin, Osana, Jones & Peterson, 2000).

1.1 Description of the Study

The goal of this research study was to design, implement, and measure the learning outcomes of a brief unit about study design and conclusions in an introductory statistics course. The research question posed in the study was: *How does introductory statistics students' conceptual understanding of study design and conclusions (in particular, unbiased estimation and establishing causation) change after participating in a learning intervention designed to promote conceptual change in these areas?*

A two-and-a-half week study design unit was designed and implemented in four sections of an undergraduate introductory statistics course at the University of Minnesota. This unit included four different activities, as well as two assessments consisting of short-answer questions: A group quiz and a lab assignment. A forced-choice assessment, the Inferences from Design Assessment (IDEA) was developed in order to be used as a pretest and posttest. All activities and assessments were reviewed multiple times by the co-advisors on this project and also by the instructors who would implement the unit. The IDEA instrument was reviewed by the co-advisors on the project and also by three statistics education experts outside of the University of Minnesota. Modifications were made to all materials based on the feedback received. The activities were also modified to be used in the online section by the researcher and instructor of the online class. In addition to the activities and assessments, lesson plans were developed for the instructors. Observation

checklist forms used during observation of the in-class sections were also developed in order to keep track of the lesson plan elements that were implemented.

The researcher met regularly with instructors prior to and during the unit in order to go over the activities and lesson plans before they were implemented, and also to review rubrics for the grading of assignments. While the unit was being implemented, the in-class sections were observed by the researcher and a graduate student co-observer, and large group discussion was videotaped. The researcher "observed" the online section by reading all discussion posts and instructor wrap-ups.

Students took the IDEA instrument just prior to the unit as a pretest, and just after the unit as a posttest. Quantitative analyses were conducted on the data from IDEA, including examination of changes in scores from pretest to posttest, and examination of response patterns for individual items. Students also took the group quiz in their randomly assigned groups, and submitted the lab assignment individually. These short-answer assessments were graded by the instructors and teaching assistants using the rubrics that were developed. As part of the data analysis for this project, the short-answer assessments were back-graded by the researcher to examine agreement in scores. A coding protocol was developed and used for qualitative data analysis of these assessments.

1.2 Structure of the Dissertation

Chapter 2 provides a review of the literature related to the learning of study design and conclusions, in particular, the purposes and roles of random sampling and random assignment. Various statistics textbooks are reviewed to examine how these topics are taught. Statistics education literature on activities and research about students' understanding of the purposes of random sampling and random assignment is presented

and summarized. Results from past statistics assessments are presented as relate to students' performance on items related to study design and conclusions. Then, literature related to conceptual knowledge and conceptual change is summarized. Chapter 2 concludes with a summary and critique of the literature presented.

Chapter 3 describes the methodology for this study. This includes the development of the activities, assessments, lesson plans, and observation forms. Chapter 3 also includes a description of the implementation of the unit and data collection. The chapter concludes with a description of the development of the coding protocol that was used in qualitative analysis of the quizzes and lab assignments.

Chapter 4 presents the results of the study, beginning with a description of the findings from the class observations. Results of the IDEA pretest and posttest are reported using descriptive and inferential methods. Finally, the results of the qualitative analysis of the group quiz and lab assignment are presented.

Chapter 5 provides a summary and discussion of the results of the study. This chapter also describes the limitations of the study, implications for the teaching of study design and conclusions in an introductory statistics course, and implications for future research. Appendices include copies of all activities and assessments, as well as full tables of analyses.

Chapter 2

Review of the Literature

The purpose of this study is to investigate students' learning of study design and conclusions, in particular the purposes and roles of random sampling and random assignment and conclusions that can be made from each. To provide background for the study, this chapter offers a review of relevant literature. First, definitions and uses of random, random sampling and random assignment are presented from the statistics literature. Then, various statistics textbooks are reviewed in order to examine how they teach and address these topics. Activities to teach about study design and conclusions that are presented in the literature are described, and research findings from the use of activities are summarized. Next, findings are presented on students' performance on items related to study design and conclusions on statistics education assessments. Literature on learning and cognition related to conceptual knowledge and conceptual change is also reviewed along with discussion of how it is relevant to students' conceptual understanding of study design and conclusions. The chapter concludes with a summary and critique of the literature that guided the development of the materials used in the unit.

2.1 Definitions and uses of random sampling and random assignment

Before addressing how the topics of random sampling and random assignment are taught, it is first important to describe how these terms have been defined in the statistics literature. This section will examine definitions of random, random sampling, and random assignment, and describe how these study designs and the scope of inferences they allow are discussed in statistical literature.

2.1.1 Defining *random*

The terms *random sampling* and *random assignment* both include the word *random*, which scholars of various disciplines have found inherently difficult to define. Randomness is an elusive concept in mathematics that does not have a precise definition (Falk & Konold, 1994). There is not a conclusive test that can establish for certain whether a particular sequence is actually random (Ayton, Hunt, & Wright, 1989). Statisticians, psychologists, and other scientists have treated randomness with ambivalence and ambiguity, and often find it easier to explain what randomness is *not* (Falk, 1991). Many attempted definitions of randomness involve complex philosophical or mathematical problems (Falk & Konold, 1997). For example, Ford (1983) and Kac (1983) write about randomness as more than unpredictability, referring to highly complex mathematical equations. Ayer (1965) writes: "what is required for the calculus of chances is a finite set of logically equal possibilities, which are fulfilled in the long run with equal frequency" (p. 49). This definition implies that randomness is governed by a probabilistic process, and has long-run predictability.

Determining whether a phenomenon is random may first involve examining the outcomes produced. Falk (1991) and Wagenaar (1991) state that randomness should be assessed by its process, not by its outcomes. For example, when looking at a specific sequence of numbers of a specific length generated by a die, it is difficult to assess whether the sequence is random because every sequence of that length is equally likely. Wagenaar argues that since humans are poor at assessing randomness of outcomes, instruction should focus on the process that generated those outcomes. Falk and Wagenaar agree that random processes have certain stable characteristics: (1) There are a fixed set of outcomes that can

happen, (2) each element selected does not depend on previous outcomes, and (3) the selection procedure does not show a systematic preference for any of the alternatives.

Since mathematicians and statisticians have had difficulty defining randomness, it is not surprising that many introductory statistics textbooks do not contain a definition of the word "random." Some textbooks do not define randomness, but use the word "random" as an adjective to modify terms such as *random phenomenon, random event, random sampling*, and *random assignment* (see for example DeVeaux, Velleman, & Bock, 2009; Moore, 2010; Triola, 2006). One statistics textbook that does define "random" is authored by Moore (2010), who writes "We call a phenomenon random if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in a large number of repetitions" (p. 263). This definition thus describes a random event as unpredictable in the short run, but showing patterns in the long run.

2.1.2 Random sampling in the statistics literature

Bellhouse (1988) and Kruskal and Mosteller (1980; 1988) provide a history of the emergence and development of random sampling methods. The importance of drawing a representative sample was first formally proposed by A.N. Kiaer in the 1890's, when the generally accepted method of collecting information was to survey an entire population. At the Berne meeting of the International Statistical Institute (ISI) in 1895, Kiaer stated that a sample based on what he called the "representative method" could provide useful information. Kruskal and Mosteller (1980) describe that the aim of Kiaer's representative method was that the sample should be "an approximate miniature of the population" (p. 175). In fact, Kiaer suggested drawing samples by lot, but never developed the idea further in his writings (Bellhouse, 1988; Kruskal & Mosteller, 1988). Rather, his idea of a

representative sample was to systematically choose districts, towns, streets, etc. that represented different social and economic conditions. He also insisted on having a substantial sample size, and stressed the importance of comparing sample demographics with as many known population demographics as possible (Kruskal & Mosteller, 1988).

Other statisticians in the early 20th century built upon Kiaer's ideas. L. von Bortkiewicz (as cited in Kruskal & Mosteller, 1980) brought up the role of probability in the representative sampling method: He raised the question of whether the observed difference between the population and sample can be considered random. Lucien March (as cited in Kruskal & Mosteller, 1980; 1988), often credited with having developed the idea of probability sampling, pointed out that randomness had been used in sampling methods to estimate the population of France. Arthur Lyon Bowley brought randomization to the forefront of survey sampling (Bellhouse, 1988). His definition of random sampling, as quoted by Neyman (1934) was: "The units which are to be included in the sample are selected at random. This method is only applicable where the circumstances make it possible to give every single unit an equal chance of inclusion in the sample." Bowley also attempted to give an empirical verification to "a type of central limit theorem for simple random sampling" (Bellhouse, 1988). He also proposed estimating parameters around the average of plus or minus three times the calculated probable sampling error. Bowley checked the representativeness of his samples by comparing his sample results to known population values, and did not find discrepancies except for two cases when he found errors in the official population statistics (Kruskal & Mosteller, 1980).

By the 1920's, two methods of sampling were considered to be standard: purposive, "representative sampling" and random sampling. Neyman (1934) provided theoretical and

practical reasons why randomization gave a more representative sample than purposive sampling. Neyman also provided alternate methods of sampling such as stratified and cluster sampling, showing that "valid" estimates of the mean were possible using these methods rather than "representative" sampling. Around the same time, Yates (as cited in Kruskal & Mosteller, 1980) explored the role of selection bias in purposive sampling, claiming that randomization helps to avoid biases resulting from personal judgment in drawing a "representative sample." Random methods of sampling have become more commonplace since the 1930's (Bellhouse, 1988).

Since then, other statisticians who have written about study design (e.g., Cornfield, 1959; Rubin, 1974) have noted the importance of being able to generalize study results to the population of interest, and random samples as the best way to ensure representativeness. In summary, the need for a representative sample has been recognized for centuries. Using randomization in the sample selection (whether it be a simple random sample or a more complex method of probability sample) is advocated in order to avoid the effects of bias.

2.1.3 Random assignment in the statistics literature

Karl Pearson and Sir Ronald Fisher's writings in the early 20th century greatly influenced modern ideas about assigning treatments to establish causality. Pearson (as cited in Cornfield, 1959) discussed a study to examine the effectiveness of a vaccine against typhoid, and noted that those who are most anxious about their health may be more likely to volunteer to receive the vaccine. Therefore, Pearson suggested inoculating "every second volunteer" in order to attempt to minimize the effect of any "spurious correlation" that may arise from the vaccinated men being more cautious about their health than the non-vaccinated men.

While Pearson addressed the importance of the two treatment groups being comparable, he did not make any mention of randomness in allocating the treatments. Fisher (1925) proposed random allocation of fertilizer treatments to agricultural plots in order to ensure that "no distinction can creep in" between pairs of plots treated alike and pairs of plots treated differently. According to Fisher, random assignment would control the probability that the treatment and the control group differ by more than a calculable amount on any variable, including those beyond the experimenters' control. Moreover, random assignment can also address the criticism: "What reason is there to think that, even if no manure had been applied, the acre which actually received it would not still have given the higher yield?" (Fisher, 1925, p. 504). Failing to randomize the assignments, according to Fisher, would overestimate or underestimate the error because pairs of plots would not provide independent pieces of information if they were systematically assigned. Cornfield (1959) later wrote that randomization controls the probability that the treatment and control group "differ by more than a calculable amount" on other variables that can influence the outcome.

Later in the 20th century, other statisticians such as White (1975), Rubin (1974), Kempthorne (1977) and Holland (1986) wrote of the advantages of using randomization in allocation of experimental trials when trying to measure causal effects of treatments. According to Holland, the "fundamental problem of causal inference" is that there is no way to know what value the response variable would take for a given subject if this subject were to undergo both the treatment and the control. In order to estimate such an effect, identical units would need to be manufactured (Kempthorne, 1977). Rubin writes that ideally, one could carry out the experiment in "matched pairs," that is, arrange the subjects

into pairs that are very similar and give the treatment to one member of the pair and the control to the other. As this is not feasible, the "impossible-to-observe" treatment effect on a single subject (or pair of matched subjects) instead gets measured as an average causal effect over a population of experimental units (Holland, 1986). Thus, randomization allows researchers to make groups comparable. Rubin showed how randomization provides an unbiased estimate of the average treatment effect. With randomization, if we have two experimental units, the response to the treatment will be the same no matter which unit receives the treatment and which unit receives the control (Rubin, 1974). Although two units cannot be identical prior to the treatment, random assignment can help make two groups comparable before treatments are applied.

In summary, random assignment is widely accepted among statisticians as the best way to ensure that treatment groups are comparable in an experiment. This mitigates the effect of confounding variables and allows researchers to determine a causal effect of the treatment.

2.2 Teaching about study design and conclusions

Introductory statistics textbooks vary in the way that they teach about the purposes of random sampling and random assignment. To examine this variation, thirteen introductory statistics textbooks and five more advanced statistics textbooks were reviewed to examine their treatment of random sampling and random assignment, and the scope of inferences one can make from each. These textbooks were chosen to represent authors that are well known in the statistics education community (e.g., Agresti & Franklin, 2009; Cobb, 1998; Moore, 2001, 2010), authors who are taking innovative approaches to teaching statistics (e.g., Lock, Lock, Lock Morgan, Lock, & Lock, 2013; Zieffler & Catalysts for

Change, 2013), and authors whose textbooks take a more traditional approach but are widely used (e.g., Triola, 2006).

2.2.1 How statistics textbooks address random sampling and generalization

To introduce the topic of sampling, many introductory statistics textbooks (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009; Devore & Peck, 2005; Lock et al., 2013; Moore, 2001, 2010; Moore & McCabe, 1999) introduce the topic of bias and convenience sampling first. For instance, a common example discussed in some of these textbooks is the 1948 U.S. Presidential election in which Dewey was incorrectly predicted to beat Truman (Agresti & Franklin, 2009; DeVeaux, et al., 2009; Lock et al., 2013; Rossman, Chance, & Lock, 2001; Utts & Heckard, 2007). This example illustrates how a sample, even though large, can provide an incorrect representation of the population because it is biased towards people with certain characteristics (e.g., wealthier voters who may vote differently from other people). Some authors such as DeVeaux et al., Moore (2001), and Lock et al. also provide the example of an Ann Landers poll that presented many negative characteristics about parenting and then asked people to write in to report whether they would still have children if they were to live life over again. This poll resulted in bias, with 90% of Americans reporting that they would not have children again. Moore and McCabe and Lock et al. contrast this example with a poll using a random sample estimating that 30% of Americans would not have children again. This example is used to illustrate how non-random samples can misrepresent the population and provide biased estimates of the true population parameter.

All 13 introductory textbooks reviewed discuss the topic of simple random sampling as the notion that every sample of size n is equally likely to be selected. Many of

the textbooks (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009; Devore & Peck, 2005; Moore, 2010; Triola, 2006; Utts & Heckard, 2007) also describe details of other types of sampling such as stratified and cluster sampling, whereas others (e.g., Lock et al., 2013; Ramsey & Schafer, 2002; Zieffler et al., 2013) briefly mention these other types of probability samples but do not go into detail. The reviewed textbooks make the basic argument that choosing a simple random sample (or any other type of probability sample) helps to obtain a sample that is representative of the population.

Introductory statistics textbooks also have different ways of illustrating the purpose of random sampling. For example, Lock et al. (2013) and DeVeaux et al. (2009) use a "pot of soup" analogy: If a pot of soup is well-mixed, then taking a spoonful of this soup will give a good representation of the taste of the soup. Moore (2001, 2010) writes that random sampling is an unbiased method because people of all characteristics (e.g., different ages, races, and socioeconomic statuses) have the same chance of being in the sample. Agresti and Franklin (2009) and Moore (2001, 2010) emphasize that allowing chance rather than human bias to select the sample will provide a more representative sample. Moreover, random sampling helps with inference because random sampling allows us to quantify the risk of a non-representative sample and also quantify sampling error (Agresti & Franklin, 2009; Devore & Peck, 2005; Ramsey & Schafer, 2002).

Some textbooks use the notion of repeated sampling to illustrate how sampling affects bias. DeVeaux et al. (2009) mention that "on average," the sample will look like the population, implying that if many samples were taken, the sample statistics would be close to the population parameter. Activity-based textbooks such as Zieffler et al. (2013) and Rossman et al. (2001) allow students to create sampling distributions based on a biased

sample and visualize how their sample statistics tend to over- or under-estimate the population parameter. Then, students create sampling distributions based on a random sample and observe how, on average, their sample statistics estimate the population parameter correctly.

Some introductory textbooks emphasize the idea that the size of the sample is not as important as the way in which it was selected. For example, Agresti and Franklin (2009), Lock et al. (2013), and Utts and Heckard (2007) all state that a randomly selected, small sample is much more useful than a poorly selected, large sample which can be heavily biased. In addition, Zieffler et al. (2013) have students actively explore this notion by quadrupling the size of the population and noting the lack of effect on the center and variability of the distribution of sample statistics.

More advanced textbooks on sampling (Levy & Lemeshow, 1999; Lohr, 2010; Thompson, 2002) also explain that random sampling avoids selection bias, but do so more formally. Similar to the introductory statistics textbooks, Lohr first introduces the topic of selection bias by contrasting the target population with the sampled population. If these two populations are different, bias occurs in sampling. Thompson explains that random sampling results in unbiasedness, which means that across all possible samples, the expected value of the estimate is equal to the value of the population parameter. Rather than referring mainly to simple random samples, Levy and Lemeshow discuss probability samples more generally as the only ones that allow reliability and validity of estimates to be evaluated from the data collected. Like the introductory textbooks, some of the more advanced textbooks also describe examples of biased samples. Lohr refers to the Literary Digest poll that incorrectly predicted Landon's win over Roosevelt in the 1936 U.S.

presidential election. Levy and Lemeshow describe how a purposeful sample to attain representation of different races may still over- or under-sample people of certain socio-economic statuses.

In summary, statistics textbooks present random sampling as a good way to avoid bias in estimation. While these books differ in their examples and illustrations, they often contrast random sampling with convenience and other non-probability methods. In all reviewed textbooks, the argument is made that random sampling will help to ensure that the sample is representative of the population.

2.2.2 How statistics textbooks address random assignment and causation

Some introductory statistics textbooks (e.g., Lock et al. 2013; Moore, 2001, 2010; Moore & McCabe, 2009; Ramsey & Schafer, 2002) introduce the topic of confounding variables before defining a randomized experiment and discussing random allocation of treatments. Other textbooks (e.g., Triola, 2006; Utts & Heckard, 2007) introduce the topic of experiments and random assignment before mentioning the topic of confounding.

Either way, many examples are given about how confounding can affect the interpretation of study results, using different contexts. For example, Moore (2001, 2010) and Rossman et al. (2001) mention the example that foreign language students tend to have a high level of English ability, but it is unclear if one variable causes the other. Triola (2006) and DeVeaux et al. (2009) give the example of a professor with a new, more stringent attendance policy who notices his attendance has gone up, but also that the better weather this year could be a confounding variable. Other textbooks use examples of headlines of studies from the media from which false causal inferences could be made. Lock et al. (2013) mention a headline claiming that hospitals have a higher risk of death

from heart attacks than casinos, and Utts and Heckard (2007) describe an observational study whose headline claimed that "prayer lowers blood pressure."

Some textbooks differentiate between lurking variables and confounding variables, mentioning that lurking variables are unmonitored variables that may be potential confounders (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009; Moore, 2001; Rossman et al., 2001; Utts & Heckard, 2007). Other books simply use the term "confounding variable" to refer to any variable that could potentially explain the association between two variables, whether the variable is measured or not (e.g., Devore & Peck, 2005; Lock et al., 2013).

The reviewed introductory statistics textbooks explain that random assignment allows for making cause-and-effect conclusions because this eliminates the effect of confounding variables. Textbooks use slightly different language to convey this idea. Many of the textbooks use the idea of "balancing" out groups so that they are similar with respect to potential confounding variables and treated alike other than with respect to the treatment variable of interest (e.g., Agresti & Franklin, 2009; Moore, 2001, 2010; Moore & McCabe, 1999; Rossman et al., 2001; Utts & Heckard, 2007; Zieffler et al., 2013). Some textbooks also use the notion of "bias" in assigning treatments, saying that random assignment prevents bias from making one treatment group different from another group (Agresti & Franklin, 2009; Devore & Peck, 2005; Moore, 2001, 2010). Other textbooks use the ideas of variation and exerting control over sources of variation. For example, DeVeaux et al. (2009) write that random assignment equalizes the effect of sources of variation that are unknown or unable to be controlled. Triola (2006) writes that random assignment provides control of the effects of variables, such that confounding does not occur.

In addition to examples of observational studies that are affected by confounding, many of the reviewed textbooks also provide examples of randomized experiments. For example, Moore (2001, 2010), Lock et al. (2013), Moore and McCabe (1999) and Utts and Heckard (2007) all describe the Physician's Health Study in the 1980's which recruited over 20,000 male physicians and randomly assigned them to take either aspirin or a placebo. When it was found that the physicians who took the aspirin were less likely to suffer from a heart attack, it could be concluded that the aspirin was the cause because random assignment made the two groups of physicians similar in all other respects. DeVeaux et al. (2009) and Moore and McCabe (1999) give an example of plants and fertilizer, where the random assignment ensures that the plots of land receiving fertilizer and the plots not receiving fertilizer are similar in all respects. In this manner, if plants do better with fertilizer, it is because of the treatment and not because of other characteristics of the soil.

Activity-based textbooks like Rossman et al. (2001) and Zieffler et al. (2013) allow students to conduct simulations to visualize how random assignment balances out confounding variables. In both textbooks, students randomly assign a number of subjects to two different treatments, first using a tactile simulation and then repeating this many times using technology. Then, students construct dotplots of aggregates of characteristics that may differ between the groups, and observe that the plots are centered at 0. Students observe that even though the groups may not be perfectly balanced in a single randomization, random assignment is a method that tends to balance out groups, on average.

More advanced statistics textbooks on experimental design also similarly explain the importance of random assignment to eliminate confounding. For example, Dean and Voss (1999) describe that experimenter bias can introduce unknown sources of variation and affect results. They give the example that a medical practitioner could assign a new drug treatment only to patients who are expected to respond well, making the drug appear effective no matter how good or bad it actually is. Similarly, Cobb (1998) discusses the problems of confounding and selection bias, using the example that students who take SAT preparation courses may do better on the exam than those who do not, simply because more motivated students are more likely to sign up for the courses. Zieffler, Harring, and Long (2011) and Cobb (1998) mention that random assignment allows researchers to attribute differences in groups to the differences in treatments, because the treatment effect of interest is isolated from other confounding factors. Wu and Hamada (2000) briefly discuss randomization as an essential component of an experiment because it mitigates the effect of variables that are not known to the experimenter but may impact the response. The basic ideas discussed regarding the importance of random assignment in planning experiments in more advanced textbooks are very similar to the ideas discussed in the introductory books.

2.2.3 How statistics textbooks make distinctions between random sampling and random assignment

While every introductory statistics textbook reviewed contained sections on sampling and experimental design, the textbooks varied in where these topics are placed. Some of the textbooks address these topics about data collection in an early chapter, after discussing the structure of data and variables (e.g., Devore & Peck, 2005; Lock et al, 2013;

Moore, 2001; Triola, 2006). Other textbooks cover descriptive statistics and exploring relationships between variables first, before mentioning data collection (e.g., Agresti & Franklin; 2009; DeVeaux et al., 2009; Moore, 2010; Moore & McCabe, 1999; Rossman et al., 2001). All of the textbooks reviewed address sampling and randomized experiments in consecutive sections or chapters, with several exceptions. Triola (2006) includes both topics in a section named "Design of Experiments." Utts and Heckard (2007) address random sampling at the beginning of a chapter about survey design, confidence intervals, and margin of error, and address random assignment at the beginning of the following chapter about experiments and examining relationships. Zieffler et al. (2013) introduce random assignment and random sampling in their second unit which involves making inferences about differences between groups, and revisit random sampling in the following unit about estimation.

Some of the textbooks reviewed provide contrasts between random assignment and random sampling, noting that the role of the randomization is different in each case. For example, Lock et al. (2013) include a paragraph describing that the role of randomness in selecting participants for a study is different from the role of randomness in assigning participants to treatments. Similarly, Zieffler et al. (2013) contrast random sampling and random assignment, outlining the purposes of each and the role of randomness in each type of study. The introductory textbook by Devore and Peck (2005) and the more advanced textbook by Zieffler et al. (2011) provide a table outlining four types of inferences that can be made: (1) generalization, (2) cause-and-effect, (3) both generalization and cause-and-effect, and (4) neither generalization nor cause-and-effect. Ramsey and Schafer (2002) have a book section on statistical inference and chance mechanisms, describing how

different randomization mechanisms allow for different scope of inference conclusions. In Ramsey and Schafer's textbook, a chart is presented describing how randomization is used for selection of units and/or allocation to groups, and what inferences these study designs allow (see Figure 2.1). Lock et al. provide a flowchart with similar information. These textbooks note that random sampling is essential for the first and third inferences noted above, and random assignment is necessary for the second and third inferences.

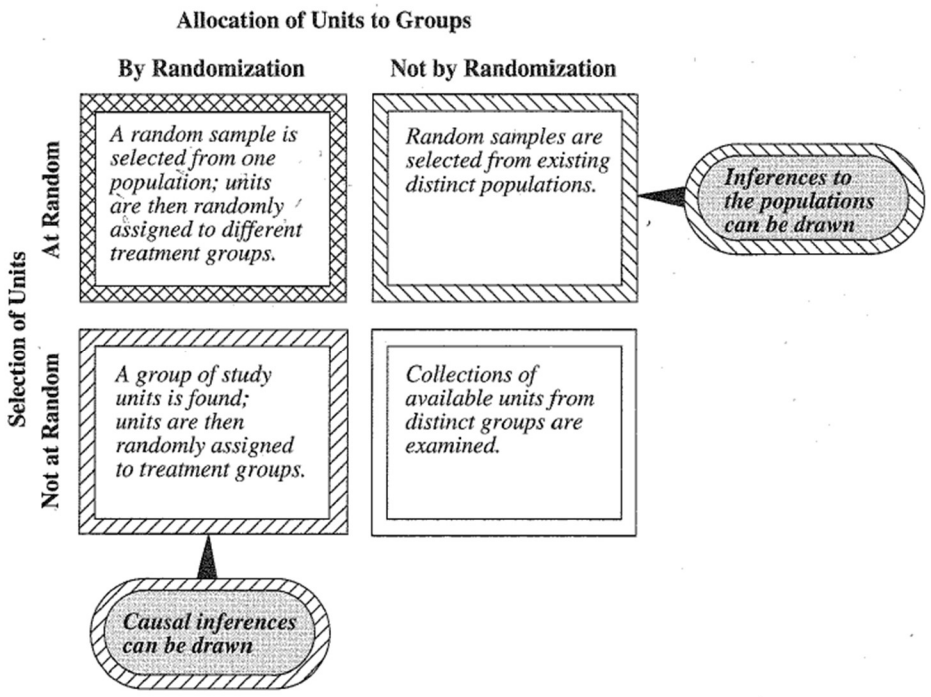


Figure 2.1. Statistical inferences permitted by study designs, from Ramsey & Schafer (2002)

Textbooks use similar ideas and language when addressing random sampling and random assignment, which may contribute to confusion between the topics. Not only do both study designs include the word *random*, but both of these study designs are helpful in reducing *bias.* For example, Agresti and Franklin (2009), Moore (2001, 2010), Moore and McCabe (1999), and Utts and Heckard (2007) discuss how bias can occur in non-random

samples, but also mention how random assignment prevents bias from making one treatment group different from another. Moore (2010) states that in studies where groups are self-selected, "personal choice will bias our results in the same way that volunteers bias the results of online polls" (p. 229). Another possible source of confusion is that some textbook authors describe random assignment as selecting a simple random sample of the participants to be assigned to each treatment. For example, Agresti and Franklin describe randomization as "pick[ing] a simple random sample of 200 of the 400 subjects." (p. 176). Triola (2006) states that "with a completely randomized experimental design, subjects are assigned to different treatment groups through the process of *random selection*" (p. 25). However, other textbooks such as Rossman et al. (2001), Thompson (2002), and Zieffler et al. (2013) use the words *random selection* to refer only to random sampling of subjects from a population, and not to random assignment of treatments.

2.2.4 Activities to teach about random sampling and generalization

Various statistics educators have published information on activities to teach about sampling. For example, Dietz (1993) describes a collaborative activity used in an introductory course in a university setting to teach methods of selecting a sample. Groups of students were asked to choose three representative samples of size 20 from a population of 317 college freshmen, and compare sample characteristics (e.g., SAT score, GPA) to population characteristics. On their own, students came up with sampling schemes such as simple random, stratified and systematic. While they successfully compared the population and sample's characteristics with regards to center, they did not consider variability. In a later modification of the exercise which included computer graphing, students were better able to compare sample plots with population plots and consider variability as well.

Like Dietz (1993), Derry et al. (2000) also created a collaborative sampling activity given in an introductory statistics course for education majors at a university. In Derry's activity, student groups were each given a large canister of colored candies, with different colors representing different majority or minority groups. After drawing repeated samples, students obtained a distribution that represented the proportion of minority candies sampled in the long run. They were given a sample of candies in an envelope and asked to judge whether this sample was "fair." Unlike Dietz's students who did not consider variability, some of Derry et al.'s students applied a two standard deviation criterion for how large the discrepancy between their observed and expected sample proportion had to be in order to label a sample as "unfair." It should be noted that Derry et al.'s activity emphasized the topic of sampling variability while Dietz's activity did not, and this may explain why Derry et al.'s students were more likely to consider variability from sample to sample.

Wagler and Wagler (2013) also designed a hands-on activity to give students experience with the process of selecting a sample for a study. Students were asked to sample Madagascar hissing cockroaches (MHCs) for their own research study which would explore whether the age of the cockroaches is a factor in food preference. Students were asked to reflect on how to select the MHCs, keeping in mind the fact that they tend to cluster in groups. This fact led students to observe that if they sampled MHCs from the same part of the container every time, they could easily end up with all MHCs of only one age group or sex. Students had to devise a plan for selecting cockroaches so that each one had an equal chance, even though it was not possible to easily number the MHCs for random selection.

2.2.5 Activities to teach about random assignment and causation

Researchers have also published descriptions of activities using hands-on work and technology to teach about random assignment in experiments. For example, Labov and Firmage (1994) had students simulate drawing a sequence of 10 random numbers by drawing pieces of paper of different sizes with these numbers. Students discussed factors that might influence the sequence of numbers drawn, such as paper size and roughness. Then, students worked with a computer program called RANDOMIZ to create random sequences of numbers between 1 and 10, and observed how each number occurs approximately the same percentage of the time. Later, students used a program called ASSIGN to randomly assign virtual plants to four treatment conditions and observed that in the long run, random assignment results in equal frequency of assignment to treatment conditions of every participant. Enders, Laurenceau, and Stuetzle (2006) also had introductory research methods students model random assignment to treatments, using a tactile simulation with a standard deck of playing cards. Students then compared the relative frequency of "background variables" (e.g., color, suit) between the groups. In a graduate level introductory research methods course, Sawilowsky (2004) used a Monte-Carlo simulation to draw repeated samples of size 4 from a large data array. The four observations were randomly assigned to one of two groups, and independent t-tests were conducted to examine differences among groups on simulated background variables. The small proportion of statistically significant t-tests across repeated trials served to demonstrate the effectiveness of random assignment to balance groups.

Statistics educators such as Derry et al. (2000) and Wagler and Wagler (2013) describe the teaching of experimental design by having beginning statistics students engage

in hands-on experiments in real-world contexts. Derry et al. asked student groups to generate hypotheses about why Wisconsin Fast Plants grown under different conditions had developed different characteristics. Student groups then designed and conducted a laboratory experiment on these plants, analyzed results and presented their findings. Derry's activity emphasized four critical components of scientifically credible evidence, one of which is that "all other competing explanatory variables (extraneous variables) must be eliminated, through randomization and control" (p. 754). As described earlier, Wagler and Wagler have students first select a random sample of Madagascar hissing cockroaches (MHCs) to examine whether age is related to food preference. Then, students are asked how to assign the roaches to the two food groups. Students reflect on various characteristics that could differ between the groups (e.g., sex, age, size) and are prompted with questions such as whether it is necessary that the two groups be identical.

2.3 Research on students' understanding of study design and conclusions

Some of the activities described above were used in research studies to explore students' understanding of the purposes of random sampling and random assignment. In addition, results from administration of statistics assessments have pointed to student difficulties understanding topics related to study design and conclusions. This subsection will discuss research findings from activities and assessments related to study design and conclusions in the statistics education literature.

2.3.1 Research results from activities that teach about study design and conclusions

Some of the previously mentioned researchers who describe the use of activities to teach about sampling and experimental design also measured student outcomes of these activities. For example, in order to examine the effectiveness of a course using

collaborative activities to stimulate complex problem solving, Derry et al. (2000) gave a pretest and posttest about various statistical concepts. For the questions on experiments and random sampling, there were significant increases from pre- to post-course. However, Derry et al. found that on assessments taken throughout the course, students had pervasive confusion about the distinction between random sampling and random assignment. This confusion also manifested itself in post-course interviews, in which students tended to over-emphasize random sampling and representativeness when it was not the most salient feature of the task at hand. Students did not bring up the lack of random assignment when it was relevant, and remained unaware that random assignment is the major experimental method for controlling sources of variation. Similarly, in an interview of high school students, Groth (2006) found that students did not bring up experimental design when it was relevant. In Groth's interviews, when students were asked for ways to determine whether a drug for the West Nile Virus was effective, most students instead came up with observational designs such as talking to doctors, and observing whether those who took the drug felt better.

Wagler and Wagler (2013) had students explore both random sampling and random assignment in designing a study to see whether age influences snack preference for Madagascar hissing cockroaches. A pretest/posttest design was used which included three questions about study design from the Assessment Resource Tools for Improving Statistical Thinking (Garfield, delMas, & Chance, 2002). For two out of three items, there was significant improvement in performance after the activity. Qualitative analysis of three additional open-ended questions given in the pretest revealed that students came in with some misconceptions about random sampling and random assignment. For example, some students preferred systematic assignment rather than random assignment because they viewed it as a better way to balance out the groups. Some students also preferred a large volunteer sample over a random sample because it obtains a wide variety of subjects. Another incorrect idea expressed was that all methods of selecting subjects are equally effective because any method works as long as there is a variety of subjects. The

set of misconceptions and reasons cited for choosing the item distractor with the corresponding misconception are shown in

Table 2.1 which appears in Wagler and Wagler (2013, p.16).

Table 2.1
Misconceptions identified by Wagler and Wagler's (2013) coding of responses to ARTIST items

Misconception	Reasons cited for choosing misconception
Misconceptions about random assignment	
• Preferring systematic assignment over random assignment	• Balances out the groups
• Preferring nonrandom assignment over random assignment	• Is a "random" way to assign groups
• All methods of random assignment are equal	• All are "random" methods; no difference between any method, all methods appropriate as long as there are equal groups
• No methods of random assignment are appropriate	• 10 samples per group are not enough
Misconceptions about random selection	
• Preferring a volunteer sample over a random sample	• Obtains a wide variety of subjects or opinions; obtains interested subjects; 200 subjects is better than 50
• Preferring a systematic sample over random sample	• Gets all possible subjects
• All methods of selecting subjects are equally effective	• Any work with a wide variety of subjects

Sawilowsky (2004) and Enders et al. (2006) implemented activities to teach random assignment and measured student learning outcomes from these activities. Sawilowsky (2004) gave a pretest to students in three sections of a graduate level introductory statistics course, asking whether they believed that random assignment of subjects could produce equal groups. The majority disagreed. One section was randomly selected to serve as the control group, reading a textbook chapter about random assignment that is similar to those found in other research textbooks. The other two sections were exposed to a Monte Carlo study which divided samples of size 4 from a virtual population into two treatment groups

of $n = 2$. Each case in the population had a "personality profile" represented by 7,500 simulated scores. Independent samples t-tests demonstrated that random assignment was successful in equalizing the two groups on 7,467 variables out of the 7,500. A posttest revealed that about three-quarters of students in the treatment group believed that random assignment can equalize groups, while fewer than 20% in the control group believed this.

Enders et al. (2006) used a 15-item multiple choice quiz to evaluate the effectiveness of an activity where students in a college-level introductory research methods class randomly assigned cards in a deck to two groups and compared the groups' characteristics. Ten of the quiz items were related to random assignment, and the others were related to other research design issues such as random selection. When comparing mean scores for the ten items dealing with random assignment, results showed that two sections of statistics courses experienced a significant increase ($p < .001$ for both sections) from pretest to posttest scores. The section consisting of introductory undergraduate statistic students showed a medium effect size ($d = .75$) and the section consisting of honors undergraduate statistics and research design students showed a large effect size ($d = .94$).

While this research suggests that using tactile and technology-based activities can increase students' understanding of issues involving random sampling and random assignment, there is not much information provided about the instruments used to assess outcomes in these studies. Sometimes, only one or a few items were used to measure outcomes (e.g., Sawilowsky, 2004; Wagler & Wagler, 2013). The next subsection reviews results of student performance on items related to study design and conclusions on larger-scale assessments.

2.3.2 Research results from the CAOS test

The Comprehensive Assessment of Outcomes in Statistics (CAOS) is an assessment with strong reliability and validity evidence that measures outcomes after a first course in statistics (delMas, Garfield, Ooms, & Chance, 2007). A sample of over 700 students from 20 institutions across the United States took the CAOS test as a pretest and a posttest. Some of the items they had the most difficulty with were related to issues of study design.

Results from implementation of CAOS revealed student problems understanding factors that allow a sample to be generalized to a population. Although there was statistically significant pretest to posttest improvement on item 38 related to random sampling and generalization, fewer than 40% of students obtained a correct answer on the posttest. Only one-fifth of the students on the pretest, and nearly 40% on the posttest, made a correct choice about the conditions that allow a generalization to the population to be made from a sample. More than 62% of the students incorrectly indicated that a random sample of 500 students would be inadequate for representing a population of 5,000 students.

Students also struggled with items related to random assignment and making cause-and-effect conclusions. Fewer than 60% of students obtained the correct answer on both pretest and posttest to two items regarding causal inference. Neither of these two items showed significant gains from pretest to posttest. For item 22, which involved understanding that correlation does not imply causation, about one-third of the students incorrectly indicated that a statistically significant correlation establishes a causal relationship between variables (despite the fact that there was no random assignment). Item

24 involved understanding that an experimental design with random assignment supports causal inferences, and just below 60% of students answered the item correctly on both pretest and posttest. The item with the worst performance on both the pretest and posttest was item 7, about understanding the purpose of randomization in an experiment (to yield treatment groups with similar characteristics). Only 8.5% of students obtained a correct response on the pretest, compared with 12.3% on the posttest, not yielding a significant learning gain from pretest to posttest. This item had the lowest pretest and posttest scores on the entire CAOS assessment. In this item, students tended to confuse random sampling with random assignment (delMas et al., 2007). Also, on the posttest, about 30% of students said that random assignment was used "to increase the accuracy of the research results," and another 30% said it was used to "reduce the amount of sampling error."

Tintle, Topliff, VanderStoep, Holmes, and Swanson (2012) also used the CAOS test. Their purpose in using CAOS was to compare students in a randomization-based curriculum with those in a consensus curriculum. On the four above items related to study design, both the randomization and consensus groups showed substantial losses in accuracy from pretest to-posttest. The loss for the consensus group was substantial, while the loss for the randomization group was minor. Tintle et al. also examined students' retention of information after the course. The three items on random assignment and causal inference (7, 22, and 24) showed better retention among the randomization cohort, while item 38 on sampling showed virtually no change.

In general, the students assessed by Tintle et al. (2012) did not perform very well on the items related to study design. The percentage correct in pretest, posttest, and retention test were under 70% for both randomization and consensus groups for item 22

(understanding that correlation does not imply causation) and item 24 (understanding that random assignment supports causal inference). Also, fewer than half of students in both groups correctly answered item 38 on the purpose of random sampling across all three test administrations. Similar to the findings by delMas et al. (2007), students had the worst performance with item 7 regarding the purpose of random assignment. On the pretest, fewer than 10% of students in each group answered the question correctly. On the posttest and retention tests, fewer than 20% answered correctly.

In summary, results from the CAOS test across different populations of students and different curricula give evidence of poor student understanding of the roles of random assignment and random sampling. Students especially struggle with being able to identify that the purpose of random assignment is to create comparable groups in each treatment.

2.3.3 Research results from the GOALS test

The *Goals and Outcomes Associated with Learning Statistics* (GOALS) test is an instrument that was originally developed to evaluate learning outcomes in a randomization-based curriculum called *Change Agents for Teaching and Learning Statistics* (CATALST: Garfield, delMas, & Zieffler, 2012). Some of the items were modified from CAOS items.

A 23-item version of GOALS was given to over 100 students in the CATALST curriculum at the University of Minnesota and North Carolina State University. For the five items related to study design and conclusions, approximately 60% of students obtained a correct answer for four of them, and just over 40% of students obtained a correct answer for one of them. When the CATALST students were compared with a national sample who had taken CAOS, the CATALST students did slightly better on two of the study design items, slightly worse on one item, and much better on one other item (Garfield et al., 2012).

A later, 27-item version of GOALS was administered to 289 students enrolled in the CATALST curriculum at six universities throughout the United States (Sabbag, 2013). The first four items on this version of GOALS relate to study design and conclusions, three of which were modified from CAOS items. The performance on these three items that were modified was generally better for the CATALST students than the performance on CAOS by the national sample of students. Similarly, Beckman, delMas, and Garfield (2017) found that students in the CATALST curriculum significantly outperformed students in a traditional introductory statistics curriculum on items regarding study design. This may not be surprising, as random sampling and random assignment play a major role in one of the units of the CATALST curriculum (Garfield et al., 2012).

The first GOALS item (modified from CAOS item 7) relates to understanding of the purpose of random assignment. About two-thirds of CATALST students answered this item correctly, much higher performance than what was seen in the CAOS sample where fewer than 15% of students answered correctly on both pretest and posttest (Sabbag, 2013). Still, 16% of students indicated that random assignment would ensure a sample that was representative of the larger population, which indicates confusion between random sampling and random assignment.

The second item on GOALS (modified from CAOS item 38) assesses the understanding of factors that allow data to be generalized to a population. On this item, 81% of CATALST students correctly indicated that a randomly selected sample of 500 students was acceptable to generalize to a population of five thousand students (Sabbag, 2013). The remaining 19% of students indicated that because of the small sample size, one could not generalize to the larger population.

The third GOALS item also measures the understanding of factors that allow data to be generalized to a population, but rather than presenting the student with a random sample, it presents a large, biased sample. Over 80% of CATALST students correctly identified that results from a call-in poll are not acceptable to make generalizations, despite the large size of the sample (Sabbag, 2013).

The fourth item from GOALS (modified from CAOS item 22) involves understanding that correlation does not imply causation. Students were asked to determine whether a strong correlation between recycling and income implies that earning more money causes more recycling. Students performed worse on this item than on the other three, with slightly less than half correctly indicating that the study design does not allow causation to be inferred. Over 20% of students indicated that one could not infer causation because of the small sample size, and another 20% indicated that the statistically significant result allows causation to be inferred (Sabbag, 2013).

In summary, although the CATALST curriculum includes random sampling and random assignment and their role in making conclusions, there is still evidence of lack of understanding of these topics. Only two-thirds of this sample was correctly able to identify the purpose of random assignment, with nearly one-fifth of the sample indicating the misunderstanding that the purpose of random assignment is to make the sample representative of the population. Also, nearly one-fifth of these students indicated the misunderstanding that a random sample that composes a small percentage of the population is inadequate for making generalizations.

2.4 Conceptual understanding of study design and conclusions

While it is possible for students to memorize that random sampling enables generalization to a population and random assignment enables cause-and-effect conclusions, it is questionable if they can apply this factual knowledge to reason effectively about study design in different contexts. Therefore, it is important to define the concepts that statistics educators deem important to learn regarding random sampling and random assignment. In order to do this, it is first helpful to examine how concepts are defined in the cognition literature.

2.4.1 Defining conceptual knowledge

There are various definitions of conceptual knowledge in the cognition literature, most often in the context of mathematics education. A key element that is present in many definitions is that conceptual knowledge involves relationships and connections among ideas. Hiebert and Lefevre (1986), who are widely cited in literature on conceptual and procedural knowledge, define conceptual knowledge as knowledge that is full of relationships, a "connected web of knowledge" where pieces of information do not stand as individual facts, but are linked to a larger network. Similarly, Tennyson and Cocchiarella (1986) define conceptual knowledge as the understanding of the structure of concepts and the relationships among them. An empirical study by Rittle-Johnson and Alibali (1999) used the definition of conceptual knowledge as "explicit or implicit understanding of the principles that govern a domain and of the interrelations between pieces of knowledge in a domain" (p. 175). Star (2005) writes that conceptual knowledge is "richly connected." In his textbook, Santrock (2011) writes that concepts group characteristics and objects based on common properties, and help learners to summarize information. All of these definitions

of conceptual knowledge include the ideas of relationships and networks between pieces of information.

Conceptual knowledge is sometimes contrasted with declarative knowledge, which involves interpreting facts about the skill domain (Anderson, 1982). More often, definitions of conceptual knowledge are given in contrast with procedural knowledge, and connections are made between them. According to Hiebert and Lefevre (1986), procedural knowledge involves understanding rules and procedures, which mainly involve sequential relations. Hiebert and Lefevre also write that concepts must be learned meaningfully, while procedures can be learned with or without meaning. Tennyson and Cocchiarella (1986) write that conceptual knowledge involves more than the storage of declarative knowledge or verbal information.

In summary, the literature suggests that conceptual knowledge involves more than just learning facts. It involves building relationships and seeing connections between ideas. Building conceptual knowledge involves both the storage and the integration of information. While conceptual and procedural knowledge are related and often grow together, the key difference between the two is that conceptual knowledge involves linking pieces of related information and not just carrying out a procedure.

2.4.2 Concepts involving random sampling and random assignment

Both random sampling and random assignment involve the notion of randomness. Prior research has found that adults have difficulty understanding random processes and reasoning about probabilistic outcomes. For example, Kahneman and Tversky (1972) reported that university undergraduates tend to judge samples as more likely if they appear to be more similar to the population, regardless of their size. Konold (1989) found that

individuals rely on context and prior knowledge to predict what will happen on the next trial, rather than considering the range of possible outcomes. Metz (1998) found that children and adults alike had difficulty reasoning about the short-term unpredictability and long-term stability of random events. Researchers have also found that adults tend to have problems identifying and constructing random sequences (Bar-Hillel & Wagenaar, 1991; Falk & Konold, 1994;1997; Olivola & Oppenheimer; 2008). Moreover, the term "random" can be problematic for students to understand. For example, many students tend to think of the colloquial definition of the word "random" as "by chance," "without order or reason," or "unexpected," rather than including the notion of probability in their answer (Kaplan, Fisher, & Rogness, 2009; Kaplan, Rogness, & Fisher, 2014). Teachers interviewed in a study by Smith and Hjalmarson (2013) similarly defined the word "random" as "out of the blue," "by chance," "unexpected," "without a pattern," and "without bias."

These difficulties understanding randomness may affect students' understanding of randomness in study design. For example, Rubin, Bruce, and Tenney (1991) analyzed interview data from senior high school students and found that students used heuristics incorrectly to reason about random sampling. For example, students tended to underestimate sampling variability, and believe that a perfectly representative sample would be obtained if it was sampled correctly. They did not properly understand the role of randomness in explaining sampling variability.

As previously discussed in section 2.2, textbooks and activities designed to teach about random sampling and random assignment often make links between the type of study design and the type of conclusions that can be made from that design. Random sampling

is linked to the ability to generalize to a population, and random assignment is linked to the ability to make cause-and-effect conclusions. Declarative knowledge, as defined by Anderson (1982) and Tennyson and Cocchiarella (1986), might involve knowing the facts that random sampling leads to generalization and random assignment leads to cause-and-effect conclusions. Procedural knowledge, as defined by Hiebert and Lefevre (1986) and Rittle-Johnson and Alibali (1999), might involve the ability to take a random sample from a population or randomly assign subjects to different treatments. In contrast, conceptual knowledge as defined by these same researchers would encompass the ability to understand why random sampling allows one to make a generalization to the population, and why random assignment can permit cause-and-effect conclusions to be made.

In order to understand random sampling and generalizations as concepts, connections should be made between the sampling method and generalization to a population. Many of the statistics textbooks reviewed (see section 2.2.1) make this connection by describing the effects of bias when a non-random sampling method is used. In order to understand the link between random sampling and generalization to a population, students may need to recognize how randomness creates a sample that is similar in characteristics to the population it represents.

Similarly, students should make connections between random assignment and the ability to make cause-and-effect conclusions. Many of the reviewed statistics textbooks (see section 2.2.2) make this link by referring to confounding variables and giving examples of how self-selection can create confounding. In order to comprehend the link between random assignment and cause-and-effect conclusions, students might need to

understand that random assignment to treatments balances out variables other than the treatment that explain any observed changes in the response variable.

Connections could also be made between random sampling and random assignment by examining the similarities and differences in the role of randomness in these two methods. In random sampling, a subset of cases is chosen at random from a larger population. Those in that subset are included in the study, and those not in the subset are excluded. With random assignment, a subset of cases in the sample is chosen at random to participate in each treatment. Thus, both random sampling and random assignment involve selection of cases at random, but the role and purpose of randomness is different in each case. With random sampling, the cases are selected randomly from the population to create a sample, whereas random assignment is done *after* the sample is chosen, selecting cases from the sample at random to put into treatment groups. Also, the issue of eliminating bias is present in both study designs. Random sampling eliminates the bias that can result in a sample being unrepresentative of the population, leading to an over- or under-estimate of the population parameter. Random assignment eliminates the bias that can cause two or more treatment groups to be different from each other in ways other than the treatment variable being assigned. The ways in which bias affects scope of inference are different, but bias is involved when there is lack of random assignment and/or random sampling.

2.4.3 Conceptual change

Research and scholarship about conceptual change may help inform how to remedy students' lack of understanding and confusion regarding the topics of random sampling and random assignment. Posner, Strike, Hewson, and Gertzog (1982) proposed a model of conceptual change that involves cognitive dissonance. This means that students first

experience some dissatisfaction with their own original beliefs, come across a new conception that is intelligible and plausible, and then revise or reconstruct those prior beliefs. Posner et al.'s theoretical model, known as the classical approach to conceptual change includes cognitive conflict as the main instructional strategy to promote conceptual change, thus requiring students to experience dissatisfaction with their current beliefs and realize the fruitfulness of new conceptions (Vosniadou, 2013).

Over the years, the idea of conceptual change has broadened beyond Posner et al.'s approach (Smetana & Bell, 2012). Research focused on constructivist and cognitive development (e.g., Vosniadou & Brewer, 1994) posits that students become aware of their existing beliefs, and then engage in activities that allow them to gradually change their conceptual structures in such a manner that they are aligned with scientifically accepted views. The framework theory approach described by Vosniadou (2012) distinguishes between *preconceptions*, which are students' initial ideas before being exposed to school science, and *misconceptions*, which are students' erroneous interpretations of the scientific concepts they learn in school science. Unlike the classical approach, the framework theory approach claims that cognitive dissonance is not necessarily required for conceptual change, and change does not happen with sudden replacement of initial conceptions after dissatisfaction is experienced. Rather, conceptual change is a slow process involving a large network of interrelated concepts (Vosniadou, 2013). It is essential to take students' prior knowledge into account, address their existing beliefs, and provide models that clarify scientific explanations (Smetana & Bell, 2012).

While much research on conceptual change has taken place in science education, these theories can also apply to students' learning of study design and conclusions in

statistics. Students' confusion between random sampling and random assignment has been documented to happen after initial instruction (e.g., Derry et al., 2000), which aligns with the idea of misconceptions developed by students after instruction, described by Vosniadou (2012). At the same time, students often come into class with erroneous ideas about study design, which may be considered incorrect "preconceptions" by Vosniadou. For example, students sometimes think that a large sample is better than a small one no matter how it was gathered (Wagler & Wagler, 2013). Application of recent conceptual change literature (e.g., Smetana & Bell, 2012; Vosniadou, 2013) to these ideas implies that it may be beneficial for students to acknowledge their beliefs and understanding of random sampling and random assignment, and gradually change this understanding so that they can distinguish between these concepts and understand how they are related to scope of inference.

Science education research suggests that the use of technology, together with an inquiry-based learning environment and guidance, can promote conceptual understanding and conceptual change (Rutten, van Joolingen, & van der Veen, 2012; Smetana & Bell, 2012). In a review of literature on the influence of technology in math education, Olive and Makar (2010) argue that technology allows students to build mathematical knowledge by bringing about a shift in empowerment from teacher as authority to students as generators of mathematical knowledge. In statistics education, technology tools, including simulations, have been used to improve students' understanding of difficult concepts such as variability, sampling distributions, and statistical inference (Biehler, Ben-Zvi, Bakker & Makar, 2013; Chance, Ben-Zvi, Garfield, & Medina, 2007). This literature implies that technology, along with an active learning environment, could be used as a tool to improve

students' conceptual understanding of random assignment, random sampling, and the scope of inferences that can be made as a result of each study design.

2.5 Discussion of the literature

Research suggests that students have trouble reasoning about randomness, random sampling, and random assignment. In this subsection, literature related to students' understanding of study design and conclusions is summarized and critiqued. Then, findings from reviews of textbooks and research studies will be discussed in order to identify possible difficulties that students may have learning about study design and conclusions, which may need to be addressed in a curriculum that teaches these topics..

2.5.1 Summary and critique

Various statistics educators have developed activities and examined their potential effectiveness in improving students' understanding of random sampling, random assignment, or both. These activities (e.g., Derry et al., 2000; Enders et al., 2006; Sawilowsky, 2004; Wagler & Wagler, 2013) generally involve hands-on and collaborative work, with or without the use of technology. The researchers who administered the activities found favorable results indicating improvement in student understanding of these concepts. However, the instruments these researchers used to assess gains in these learning areas have generally contained only a few items, and have also lacked evidence of psychometric strength. Enders et al. (2006) do not provide any information regarding the reliability or validity of the 15-item quiz used to measure outcomes in their study. Although Wagler and Wagler (2013) used items from ARTIST which is a high-quality assessment developed by many teachers and researchers (Garfield & delMas, 2010), there was significant improvement in only two out of three items used. Sawilowsky (2004) did not

use an assessment of cognitive outcomes, but rather used only one item to examine whether students believed that random assignment was effective in balancing out confounding variables. Derry et al. (2000) gave a pretest and posttest covering a variety of statistical topics to evaluate the curriculum, but also did not provide detailed information regarding its reliability or validity evidence. However, one strength of Derry et al.'s study is that interviews were conducted and then scored independently by two researchers who then met to resolve any discrepancies. In order to have more convincing evidence of the effectiveness of activities to help understanding of study design and conclusions, more psychometrically strong instruments are needed to assess understanding of these areas.

The assessments used in the above research studies were given to students at different time periods. Enders et al. (2006), Sawilowsky (2004), and Wagler and Wagler (2013) assessed students prior to and immediately following the learning activities they administered. Derry et al. (2000) developed an entire curriculum and gave a larger assessment (including many topics other than study design) before and after the course, in addition to conducting post-course interviews. While students could develop better understanding immediately after an activity (as was found by Enders et al., Sawilowsky, and Wagler and Wagler), this does not mean that the better understanding will be retained after the course is finished. Thus, in order to determine the effectiveness of learning interventions dealing with random sampling and random assignment, it would be valuable to measure learning retention rather than only measuring understanding immediately after the activity.

Although the use of psychometrically strong assessments such as CAOS (delMas et al., 2007) as a course pretest and posttest has revealed student difficulty with items

related to study design, these types of assessments have not been used to test the effectiveness of an educational intervention designed to teach about the purposes of random sampling and random assignment. Findings from administration of the CAOS test (delMas et al., 2007) were from a national sample including a wide variety of curricula. The administration of CAOS by Tintle et al. (2012) was done as a pretest, posttest, and retention test across two different types of curricula (simulation-based and consensus) but Tintle et al. do not go into depth on how concepts of generalization and causation were taught in their curricula.

Literature on understanding concepts and conceptual change was also reviewed. Much of the literature on conceptual change in science education deals with erroneous beliefs that students bring in with them before the course, and that may persist even after instruction. Beliefs that students hold prior to class exposure to topics involving study design may be called "preconceptions" as defined by the framework theory to conceptual change (Vosniadou, 2012). As discussed earlier, introductory statistics students may have incorrect preconceptions that systematic assignment is better than random assignment, or that larger samples are always better than smaller samples regardless of collection method (e.g., Wagler & Wagler, 2013). Also, erroneous beliefs can develop as a result of instruction. These erroneous interpretations of concepts that students learn may be called "misconceptions" as defined by the framework theory to conceptual change (Vosniadou, 2012). For example, students may learn about random assignment and random sampling, but develop problems distinguishing between the two study designs and the types of conclusions supported by each method (Derry et al., 2000).

2.5.2 Possible difficulties in understanding concepts related to study design and conclusions

In designing a curriculum to teach about study design and conclusions, it is useful to consider potential difficulties students may have in learning about study design and scope of inferences. Many statistics textbooks and courses address proper methods of data collection and the scope of inferences that can be made from different study designs. Thirteen introductory statistics textbooks were reviewed, which all included discussion of the use of random sampling to make generalizations to a population, and of the use of random assignment to enable cause-and-effect conclusions. While these fundamental ideas were included in textbooks, there was much variation in the order and way in which they were presented, which may shed light on why random sampling and random assignment can be confusing concepts to learn.

For example, most of the textbooks reviewed included random sampling and random assignment in the same section or chapter, often sequentially. While this makes sense given that they both pertain to methods of data collection, this proximity in location within the textbook and curriculum may cause the two concepts to blur together. Students may learn that it is good practice to use random mechanisms in study design, but could find it more difficult to understand exactly how randomness is used to make different types of conclusions. Some books make connections between random sampling and random assignment, comparing and contrasting the different roles that randomness plays in the design and the scope of inferences this allows (e.g., Lock et al., 2013; Ramsey & Schafer, 2002). However, many other textbooks teach these concepts in separate sections, without explicitly comparing how random sampling and random assignment are similar or

different. Without having to think about these comparisons, it may be easier for students to confuse these two similar, yet distinct concepts related to randomness in data collection.

The similar vocabulary used to teach about random assignment and random sampling may also contribute to students' inability to distinguish between the two. For example, they both contain the word "random" and both involve random selection to separate some units from others. In the case of random sampling, the random selection separates the units included in the study from the units not included in the study. With random assignment, the random selection separates the sample units included in one treatment from the units included in another treatment. In both cases, the randomness eliminates bias. With lack of random sampling, bias results in the sample being systematically different from the population. With lack of random assignment, bias results in groups being systematically different from each other with respect to possible confounding variables. Some textbooks use the term "bias" and describe the way bias is mitigated by "random selection" when discussing both random sampling and random assignment (e.g., Agresti & Franklin, 2009; Triola, 2006). While these uses of the terms "bias" and "random selection" are accurate and point to similarities between random sampling and random assignment, they may contribute to students' failure to distinguish between the two.

One possible reason that the purposes of random assignment and random sampling are difficult to understand is that they are not merely facts to be memorized but arguably concepts to be learned. Conceptual knowledge is full of connections and relationships (Hiebert & Lefevre, 1986). These concepts are full of connections between study design and conclusions to be made. Random sampling is connected to generalization, because the

randomness helps to mitigate the effects of systematic bias in selecting units from a population, thus helping to ensure representativeness. Random assignment is connected to cause-and-effect conclusions, because the randomness helps to balance out confounding variables that may provide alternative explanations for associations found between the explanatory and response variable. While students could merely memorize that one can generalize to a population with a random sample and one can make cause-and-effect conclusions with a randomized experiment, this is arguably not conceptual understanding. In order to understand these two study designs as concepts, one has to make a link between random sampling and generalization, and random assignment and cause-and-effect conclusions.

In the literature reviewed, there is evidence that students come in with incorrect preconceptions that can get in the way of their understanding of concepts related to study design and conclusions, and can also develop misconceptions as they learn. For example, Sawilosky (2004) found that students came into an introductory statistics course with the disbelief that random assignment balances out groups with respect to confounding variables. Wagler and Wagler (2013) identified a series of misconceptions using qualitative data analysis of a pretest (See

Table 2.1 in section 2.3.1). Some of these erroneous ideas include that systematic ways to choose a sample are preferable than random sampling, and that sample size is more important than method of sample selection. Students may also have difficulty distinguishing between the purposes of random sampling and of random assignment, as found by Derry et al (2000) in post-course interviews.

Some researchers have noted that even after completing an introductory statistics course, students have trouble reasoning about these ideas. On the CAOS and GOALS assessments, students had particular difficulty with items related to study design and conclusions, even when these assessments were administered across different student

populations. Derry et al.'s (2000) research revealed that students tended to mix up the topics of random sampling and random assignment. In this study, post-course interviews revealed that students had trouble making distinctions between these two study designs.

Quantitative data from assessments such as CAOS and GOALS can help point to false understandings that students have, but can only give insight based on specific distractors that students chose on items. For example, students may emphasize sample size over sampling method and think that a sample must be sufficiently large (especially relative to the population) in order to make generalizations to a population (e.g., CAOS item 38; delMas et al., 2007). Students may also believe that a strong enough, statistically significant correlation is enough to make a causal claim (e.g., CAOS item 22), or that the purpose of random assignment is to increase accuracy of research results or reduce sampling error (e.g., CAOS item 7).

In summary, recognizing the difficulties that students can have in understanding concepts related to study design and conclusions may be helpful in designing curriculum materials to target their incorrect ideas. Recognizing these difficulties can also be helpful for designing assessment items with distractors to detect specific misunderstandings.

2.5.3 Problem statement

Although statistics education recommendations for students in introductory statistics courses include learning about the role of randomness in study design and conclusions, past research suggests that this is not easy for students to learn. Activities to learn about random sampling and/or random assignment have been developed and some learning outcomes from these activities have been measured. Also, results from large-scale

assessments have revealed information about students' understanding of topics related to study design and conclusions in different course curricula. However, no published study exists to date involving the implementation of a unit specifically on study design and conclusions, or attempting to measure the effectiveness of such a unit. Therefore, the aim of this study is to develop a study design unit to be implemented in an introductory statistics course, and assess students' understanding of concepts related to study design and conclusions.

Chapter 3

Methods

3.1 Introduction

The research question that this study attempted to answer is: *How does introductory statistics students' conceptual understanding of study design and conclusions (in particular, unbiased estimation and establishing causation) change after participating in a learning intervention designed to promote conceptual change in these areas?*

This chapter begins with an overview of the study and a description of the course in which the study design unit was implemented. The development of the course activities, assessments, and all accompanying materials (e.g., lesson plans, assessment rubrics) is described. The chapter then provides information about how the study design unit was implemented. Finally, data analysis methods are described, including the development of a coding scheme to aid in qualitative analysis of open-ended assignments.

3.2 Overview of the study

To answer the research question stated above, a two-and-a-half-week study design unit was developed to use in an introductory statistics course. The study took place during the spring semester of 2016. The unit was implemented in four sections (three in-class, one online) of a one-semester undergraduate three-credit introductory statistics course (EPSY 3264, Basic and Applied Statistics) offered by the Department of Educational Psychology at the University of Minnesota. The study design unit included four activities, one group quiz, and one lab (homework) assignment. Lesson plans for instructors were developed for each of the four activities. All of these materials were reviewed by two faculty members at the University of Minnesota, Dr. Robert delMas and Dr. Andrew Zieffler, co-advisors on

this project. The activities, quiz and lab assignments were also reviewed by the three instructors of EPSY 3264, and modifications were made based on all of these reviews. The study design unit was implemented by the three instructors of the four sections of EPSY 3624 (with one instructor teaching two sections). Instructors met regularly with the researcher prior to the implementation of the activities. During the study design unit, the researcher observed class sessions along with a co-observer for the in-class sections. The researcher read all online discussions and group discussion summaries for the online section.

The forced-choice Inferences from Design Assessment (IDEA) was developed to assess students' conceptual understanding of study design and conclusions before and after the curriculum. This assessment was first reviewed by the project co-advisors Dr. Robert delMas and Dr. Andrew Zieffler, and was then also reviewed by three statistics education experts at other institutions. Modifications to IDEA were made based on all of these reviews. Students completed the IDEA online prior to the start of the study design unit as a pretest, and again as a posttest at the completion of the study design unit. The student responses to the pretest and posttest were analyzed quantitatively, and students' constructed responses to the group quiz and lab assignment were analyzed qualitatively.

3.3 Course and participants

The curriculum was implemented in an undergraduate introductory statistics course. The researcher had taught this course before, but was not involved in teaching or assisting with the course during this semester. This introductory statistics course is commonly referred to as the Change Agents in Teaching and Learning Statistics (CATALST) course (Garfield et al., 2012). This is an innovative course originally

developed by researchers at the University of Minnesota, using a simulation-based approach to teach the ideas of inference. The pedagogical approach for each lesson in this course was to have students spend most of class time discovering concepts by cooperatively working on activities. After each activity, the instructor led a large group discussion to wrap up the main ideas in the activity. The activities for this study design unit were developed with this pedagogical structure in mind.

3.3.1 Class sections and teaching staff

The curriculum was implemented in all four sections of the course. There were three in-class sections. Section 1 met on Mondays and Wednesdays from 1:00-2:15pm, Section 2 met on Tuesdays and Thursdays from 9:45-11:00am, and Section 3 met on Tuesdays and Thursdays from 1:00-2:15pm. Sections 1 and 3 were taught by the same instructor, and section 2 was taught by another instructor. Both of these instructors were PhD candidates focusing on statistics education. They had been teaching the CATALST course for at least two years and had worked on revising the course each semester. There was also a fully online section of the course (section 4) taught using the course management system Moodle. The instructor of the online course was a PhD student in statistics education who had taught and revised the CATALST course previously, and was also an experienced high school statistics teacher. The online course used the same basic activities (modified to focus on several key questions for students to post answers on discussion boards), and also used the same assessments. Each section had one teaching assistant. Sections 1 and 3 had the same teaching assistant who typically attended class once per week. Section 2 had another teaching assistant who also attended class regularly, though

not every day. Section 4, the online section, had a third teaching assistant who helped with grading of assignments.

3.3.2 Students

The participants who experienced the study design unit during the spring semester of 2016 were undergraduate students at the University of Minnesota taking the CATALST course. Most students took this course to fulfill a mathematical thinking general education requirement. Students were of many different majors, most of which did not heavily involve mathematics.

One week prior to the start of the curriculum, the researcher visited all three in-class sections for 5-10 minutes to briefly explain the purpose of the study and what would happen. For the online class, an e-mail was sent with this information (see Appendix A1). A video made by the researcher explaining this information was shared with the online class. For the in-class sections, the researcher explained to students the same information that was in the e-mail, except that for in the in-class sections, the class would be observed by the researcher and a co-observer. The researcher also explained that the class would be videotaped, with the camera pointed at the instructor. Students were asked for their permission to use their de-identified responses to the IDEA pretest and posttest, Group Quiz 5, and Lab 8, which were the assignments for this curriculum. Students received a consent form explaining this (see Appendix A2), and were allowed to opt out of participating in the research study. A total of two students opted out, one from section 1 and one from section 3.

3.3.3 Class observers

In order to provide extra observations, two graduate students in statistics education agreed to attend the in-class sections along with the researcher in order to take notes on how the curriculum was implemented and how students participated. Both co-observers were PhD students in statistics education at the University of Minnesota, and both had experience either teaching or being a teaching assistant for the CATALST course. One of the observers attended both sections 1 and 2, and the other observer attended section 3. There were no co-observers for the online class, as the discussion for that section occurred entirely through online discussion boards and Google Docs. For this online section, the researcher could view the entirety of all groups' discussions, so there was no need for a co-observer.

Prior to each activity, the observers were given a copy of the class activity, the corresponding lesson plan given to the instructors, and a corresponding observation form. Both the researcher and the co-observer used these forms for the class observations. The observation forms (see Appendix E) contained a checklist with the elements of the lesson plan, so that observers could check off components that were implemented. For example, each discussion question on the lesson plan had a checkbox next to it, and whenever that question was asked by the instructor, the observers checked it off. Also, the form contained potential issues for students anticipated by the researcher, which the observers then checked off if they saw students encountering any of these issues. There was additional space to take notes for each part of the activity. Observers were also asked to consider in general what students seemed to be understanding well, where students seemed to be struggling, and how the instructor was dealing with student questions. The structure of the

observation forms and the process for observing the classes was discussed with the observers in a meeting with each of them one week prior to the start of the unit. For the group quiz, there was no lesson plan as there were no activities to implement. Rather than receiving an observation form, the observers simply received a copy of the group quiz and were asked to observe how students discussed their reasoning to answer the questions, what they seemed to be understanding well, and what difficulties they seemed to have.

3.4 Development of activities

Before the curriculum was developed, two learning trajectories were developed, one for helping students to learn about random sampling and generalization to a population, and one for helping students to learn about random assignment and establishing causation. The learning trajectories were developed based on reviews of statistics textbooks that teach these topics (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009; Devore & Peck, 2005; Lock et al., 2012; Moore, 2001, 2010; Moore & McCabe, 1999) as well as reviews of literature related to students' understanding of study design and scope of inferences (e.g., Derry et al., 2000; Sawilowsy et al., 2004; Wagler & Wagler, 2013).

Then, it was determined that three activities were needed: one to teach about sampling methods, unbiased estimation and generalization to a population; one to teach about methods of assignment to groups in an experiment and establishing causation; and one to help students distinguish between random sampling and random assignment. The CATALST curriculum incorporated group quizzes every few days. Therefore, a group quiz was added to the schedule, not only to fit with the structure of the curriculum, but also to provide qualitative data. Also, one of the existing course activities ("Murderous Nurse") was added because fit well with the curriculum, as it gave students a context in which there

are limitations to scope of inferences, when neither random sampling nor random assignment are possible.

A five-day study design curriculum was designed for the in-class sections as shown in Table 3.1. For the online class, a total of three weeks was spent on this curriculum: During the first week, the "Sampling Countries" activity was completed. During the second week, the "Strength Shoe" and "Murderous Nurse" activities were completed. During the third week, students completed the "Survey Incentives" activity and then Group Quiz 5.

Table 3.1
Study design curriculum

Day	Topic	Activity Name	Reading prior to activity
1	Sampling methods and unbiased estimation	Sampling Countries	None
2	Assignment to experimental groups and establishing causation	Strength Shoe	Establishing Causation
3	Observational studies	Murderous Nurse	Scope of Inferences
4	Study design and scope of inference	Group Quiz 5	None
5	Distinguishing between random sampling/generalization and random assignment/establishing causation	Survey Incentives	None

Two readings were developed. One reading called "Establishing Causation" (Appendix B2), to be read before the "Strength Shoe" activity, introduced students to the ideas of observational studies, confounding variables, random assignment, and cause-and-effect conclusions. Another reading called "Scope of Inferences" (Appendix B4) summarized the two types of conclusions students had learned about (generalizing to a

population and establishing causation) and distinguished between the two types of randomness necessary to make each conclusion (random sampling and random assignment, respectively). The "Scope of Inferences" reading was to be done after students had learned about random sampling and random assignment, but before the "Murderous Nurse" activity in which they examined an observational study that had no random sampling. There was no reading about random sampling before the "Sampling Countries" activity. This decision was made to ensure that students would complete the IDEA as a pretest without having had prior exposure to any part of this unit.

Two short-answer assessments were developed as a part of this unit: a group quiz and a lab (homework) assignment. In the CATALST course, it was customary to have a group quiz every few class days and also to periodically collect individual lab assignments. In-class instructors randomly assigned students into groups of two or three students, and students completed activities and group quizzes in these assigned groups. Online, the instructor randomly assigned groups of 4-6 students, and students completed group quizzes in GoogleDocs online. The in-class instructors chose to change the groups every few weeks, but the online instructor chose to keep groups the same throughout the semester. The quiz occurred during the unit, and the lab assignment was due after the completion of all activities in this unit. Both the quiz and the lab assignment were created to assess students' understanding of study design and conclusions, with the goal of having them apply their knowledge of study design and conclusions to real or realistic studies in various contexts.

3.4.1 Order of activities

The order of the activities in the study design unit was carefully considered, as it is possible to teach about causation first and generalization second, or the other way around. The introductory statistics textbooks that were reviewed prior to the development of the unit varied greatly in the order and placement of study design topics. Most of the textbooks reviewed (e.g., Agresti & Franklin, 2009; Devore & Peck, 2005; Lock et al., 2013; Moore, 2001, 2010; Rossman, Chance & Lock, 2006; Utts & Heckard, 2007) presented sampling issues and generalization first before addressing random assignment and causation. However, some of the textbooks reviewed (e.g., Moore & McCabe, 1999; Zieffler & Catalysts for Change, 2013) addressed experimental design and causation before addressing sampling and generalization. After a review of the limited literature about statistics students' understanding of study design and conclusions, no evidence was found as to whether it would be more beneficial to introduce sampling and generalization first, or experimental design and causation first. In statistical studies, the first stage is to choose a sample of participants, and any potential assignment to groups happens only after the initial sample is chosen. In order to reflect this natural order of data collection, the decision was made to place the "Sampling Countries" activity about sampling and unbiased estimation before introducing the "Strength Shoe" activity about random assignment and causation.

For the three in-class sections, the group quiz was placed between the "Murderous Nurse" and "Survey Incentive" activities. This was done because it was customary in the CATALST course to spread out assessments rather than cluster them together. Therefore, it was not desirable to assign the group quiz and the lab assignment on consecutive class days at the very end of the unit. Before taking the group quiz, students had already learned

about sampling and unbiased estimation, and had learned about random assignment and causation. They had also read the "Scope of Inferences" reading which summarized and distinguished between these two types of study design and conclusions.

Due to the structure and timeline of the online course, it was not possible to place the group quiz in between the "Murderous Nurse" and "Survey Incentives" activity as was done with the in-class sections. This is because of the additional time it takes for students to complete a group quiz in an online asynchronous environment. In the online course, students completed the "Survey Incentives" activity and Group Quiz during the same week. The summary for the activity was due two days before the group quiz. In the online course, discussion summary deadlines were typically on Wednesdays and group quizzes were typically due on Fridays, so this same schedule was maintained for this activity and quiz as well.

3.4.2 Development of course readings

Two readings were written as a part of this curriculum to be assigned to be completed before certain class days. There was no reading about sampling before the "Sampling Countries" activity so that students would be sure to complete the pretest before experiencing the readings and activities in this curriculum. However, as the concepts of random assignment and ability to make cause-and-effect conclusions are complex and required the introduction of some terminology (e.g., "explanatory" and "response" variables, "confounding"), a reading called "Establishing Causation" was drafted (see Appendix B2). This reading defined explanatory and response variables, distinguished between association and causation, and defined confounding variables.

A second reading called "Scope of Inferences" (see Appendix B4) was drafted in order to help contrast the two types of study design, random sampling and random assignment, and distinguish between the types of conclusions that could be made from each. This reading was presented to the students after they had been introduced to the concepts of sampling and generalization, and assignment to groups and causation, through the "Sampling Countries" and "Strength Shoe" activities. Using the example of the Physician's Health Study (http://phs.bwh.harvard.edu/), generalization, sampling and bias were first discussed, including the implications of the fact that the sample for this particular study consisted of recruited male physicians ages 40 to 84. In the next section of the reading, concepts of causation, confounding, and random assignment were discussed. The reading discussed how physicians were randomly assigned to take aspirin or placebo in the Physician's Health Study and how confounding variables such as health habits tend to balance out so that any differences in heart attack rate can be attributed to the aspirin treatment. In the last section of the reading, the differences between generalizing to a population and establishing causation were highlighted using this context. The reading explained how a study can have random assignment, random sampling, both, or neither, and it is difficult realistically to have both.

After the readings were drafted, they were sent to the project advisors for feedback and some changes were made to clarify wording and increase consistency in terminology (e.g., using the word "participants" rather than alternating between "participants" and "subjects"). Also, suggestions were made to summarize the two types of study design and scope of inferences in a table. Table 3.2 below was shown near the end of the "Scope of Inferences" reading. The readings were then sent to the instructional team, who suggested

that the readings were too dense compared to other readings that students in the CATALST course that semester had been encountered so far. Therefore, the readings were shortened as much as possible without losing important concepts. Longer sentences and paragraphs were broken up into shorter ones, and the language was simplified. The text of the readings was identical for the in-class and online students. For the in-class students, the readings were shared as documents separate from the activity documents, whereas for the online students, the readings were integrated into the activity file. This was done so that the online students would not miss the readings before starting their activities.

Table 3.2
Table contrasting random sampling and random assignment shown in "Scope of Inferences" reading

		Selection of Units	
		Random Sampling	**No Random Sampling**
Allocation of Units to Groups	**Random Assignment**	Can make a causal conclusion and can generalize conclusion to the population.	Can make a causal conclusion but cannot generalize this conclusion to the population
	No Random Assignment	Can generalize to the population, but cannot make causal claims.	Cannot generalize to the population, and cannot make causal claims either.

3.4.3 *Sampling Countries* activity

The first activity in this curriculum was intended to develop conceptual understanding of sampling and bias, and specifically why random sampling is an unbiased sampling method. This activity was inspired by a previous activity called "Sampling" from the CATALST curriculum (Zieffler & Catalysts for Change, 2015, pp. 162-176). In this activity, students worked with a "population" of all words from the Gettysburg address, sampling ten words which they considered to be "representative of the passage" and contrasting the mean word length from their convenience samples with the mean word lengths plotted from random samples. The new activity "Sampling Countries" incorporated this notion of contrasting students' self-selected sample estimates with random sample estimates, but with some major differences from the previous "Sampling" activity. First, it was determined that students needed to learn with a more meaningful and realistic context, in accordance with recommendations from literature on cognition and conceptual change (e.g., Bransford, Brown, & Cocking, 2000; Vosniadou, 2013). At the same time, a context was needed where a population was accessible from which to sample. The context in the "Sampling Countries" activity involves sampling countries from the population of countries of the world in order to estimate their average life expectancy, using data from the World Bank (http://www.worldbank.org). Life expectancy was chosen because it was a variable that could be of interest to the students, while at the same time not containing missing data for a large number of the countries. (For example, one variable with values available for all countries was land area, but this variable was considered less likely to capture students' interest.) While there were a few countries that did not have data on life expectancy, there were 196 countries that did have available data, and in the activity these

were considered to be the "population." Another reason for choosing life expectancy as a variable to focus on is that it was anticipated that the students (most of whom are from the United States) would tend to more easily recall countries with higher life expectancies, thus finding that their convenience sampling would be a biased method.

Another major change from the original "Sampling" activity is that the "Sampling Countries" activity would focus more deeply on fewer concepts. This is in line with the recommendations from the conceptual change literature on emphasizing depth over breadth (e.g., Vosniadou et al., 2011). For example, the original "Sampling" activity addressed not only issues of biased and unbiased sampling methods, but also concepts of sample size and variability, and the idea that population size does not affect the tendency of random sampling to produce unbiased estimates. Instead, the "Sampling Countries" activity focused mainly on contrasting biased with unbiased sampling methods (which are ideas present in nearly all of the textbooks reviewed previously). One of the most common misconceptions that has been documented regarding student's reasoning about sampling is that larger samples are always better, regardless of sampling method (delMas et al., 2007; Sabbag, 2013; Wagler & Wagler, 2013). Therefore, "Sampling Countries" was designed to address this misconception rather than focusing on sample size and variability of sample statistics, an idea which would be seen later on in the course.

***"Sampling Countries" initial draft*:** The initial draft of the "Sampling Countries" activity began with an introductory reading about unbiased estimation, including some examples from real-world studies where sampling went wrong. This reading occurred as part of the activity because students had to complete a pretest before this class, and it was not ideal for them to have a reading due on the same day as the pretest. After the initial

reading, data were presented from the World Bank in the activity. The activity asked students to explore this population by plotting all of the countries' life expectancies, and also find the percentage of countries that had more than half of the population living in urban areas. Both a categorical and quantitative variable were included so that students could see the same concept at work from multiple perspectives, following recommendations from cognition literature to provide many examples in which the same concept is at work (Bransford et al., 2000; Donovan & Bransford, 2005). The activity then continued in three major parts:

(1) A set of samples of 10 countries each were shown, each recalled by a hypothetical student. Students were asked to examine sample statistics from these samples and determine whether or not this method of convenience sampling tended to over- or under-estimate the true parameter.

(2) The activity then presented similar plots of sample statistics from convenience samples with a larger sample size (n=25), and asked students to determine whether or not taking a larger convenience sample mitigated bias. This part of the activity was meant to address the misconception that larger samples are always better, regardless of sampling method.

(3) The activity asked students to use *TinkerPlots*^{TM} to take many random samples of size 25, plot the sample statistics for each of the two variables, and determine whether random sampling was an unbiased sampling method based on where each plot was centered. The activity then concluded with a brief summary reading of biased vs. unbiased sampling. The reading mentioned that in real life we don't actually know the population parameter and we only have one sample, so it is important to use an unbiased sampling

method like random sampling in order to be able to use the sample to generalize to a population.

"Sampling Countries" revisions: The "Sampling Countries" activity went through various rounds of feedback. First, it was reviewed by the two co-advisors on this project, and changes were made based on this feedback in order to include more active learning. For example, originally, the convenience samples had been given to the students to save time. Instead, one of the major changes made based on this first round of feedback was that students were asked to come up with their own samples of size 10 and graph their sample statistics along with the rest of the class's estimates. Then, students were asked to simulate drawing biased samples of size 25 using a *TinkerPlots*^{TM} sampler created for this purpose. This was intended to allow students to have an active role in the convenience sampling portion of the activity, although it would take more time than being given the samples.

After these initial modifications, the activity was sent to the instructors who would implement it for feedback. Based on this feedback, the activity went through additional major changes in order to align more with typical activities in this course. First, it was judged to be far too long for the students in the course to complete. Therefore, much of the reading was cut and some was replaced with brief discussion questions to address the same concepts. It was estimated that taking two sets of biased samples (one smaller and one larger) would take up too much class time. Therefore, in order to address the misconception that larger samples are always better regardless of sampling method, the activity was changed so that students would first take a convenience sample of size $n = 20$ and then random samples of size $n = 10$. In this manner, the activity could still help students to discover the idea that a smaller random sample is better than a larger biased one. Moreover,

the categorical variable (whether or not more than half of the country's population was urban) was cut in the interest of time and emphasizing depth over breadth, so that students could focus only on the life expectancy variable.

After more feedback from the instructional team and co-advisors, the activity went through more wording and formatting changes in order to be more consistent with the existing activities in the course. The amount of reading was drastically shortened and some of the points that had been addressed in the activity readings were instead addressed in additional activity questions or in the wrap-up questions given to the instructors in the lesson plan. Also, some questions that seemed to address the same concepts as other questions were cut and longer questions were split into smaller sets of questions. Some visuals were added near the beginning of the activity to introduce the concept of biased and unbiased sampling methods.

The final "Sampling Countries" activity had students contrast taking convenience samples of size 20 with taking random samples of size 10, and comparing the plots of sample mean life expectancies using each method to determine whether each method is biased or unbiased. They also considered whether a smaller random sample is better than a larger sample of countries recalled by classmates. The final version of the "Sampling Countries" administered to the three in-class sections can be found in Appendix B1.

3.4.4 *Strength Shoe* activity

The second activity in this curriculum was designed to help students' conceptual understanding of how random assignment helps to balance out confounding variables in the long run. The "Strength Shoe" activity already existed in the previous course curriculum (Zieffler & Catalysts for Change, 2015, p. 147-158). In this activity, students

were presented with the context of determining whether a particular shoe called the Strength Shoe increases jumping distance over ordinary athletic shoes. Students used *TinkerPlots*^{TM} to randomly assign twelve participants into each of two groups: one group wearing the Strength Shoes and one group wearing the ordinary shoes. They repeated this randomization many times and first plotted differences in average heights and differences in percent of females in each group (two confounding variables that have been observed for each subject). Then, they revealed a hidden "unobserved" confounding variable called the "X-factor" and also plotted the differences in averages for this variable. The activity was intended for students to observe that these differences were all centered at 0, thus indicating that random assignment tends to balance out these confounding variables in the long run, even though the two groups may not be exactly equal in a single randomization.

"***Strength Shoe" initial revisions:*** Because this activity used a realistic and meaningful context, and already addressed the concepts of confounding, random assignment, and balancing out confounding variables, it was taken and modified for this curriculum. The sample size in the activity (n = 12) was kept small in order to address students' potential misconception that random assignment does not work at all with small samples (Sawilowsky, 2004). The activity had students reveal the hidden "X-factor" and plot the differences in percent of subjects with the X-factor for many randomizations. The part of the activity that asked them to plot differences in percent females was cut in order to save time.

One major addition was made to the activity to address the misconception that purposeful assignment is better than random assignment for balancing out confounding variables. As documented by Wagler and Wagler (2013) and Sawilowsky (2004), students

tend to be skeptical of the ability of random assignment to balance out confounding variables. Instead, students may believe that it is better for humans to purposefully assign the groups in order to balance out as many confounding variables as possible. Therefore, a part of the activity was added near the beginning where students were presented with a purposeful assignment of two groups with an even number of males and females in each group, and a similar average height. Students were then asked to reveal a hidden genetic "X-factor" and notice that the Strength Shoe group had a much higher percent of subjects with the "X-factor" than the ordinary athletic shoe group. Then, students went on to randomly assign the twelve subjects into two groups using *TinkerPlots*^{TM} and plot the difference in average heights for many randomizations.

The activity was designed so that students would observe that even though a single random assignment does not perfectly balance out groups with respect to all confounding variables, on average, across many random assignments, both known and unknown confounding variables balance out. Questions were added at the end of the activity that asked students to answer questions about the ability to make causal claims from random assignment, and to revisit the idea of why it is important to consider potential effects of unobserved confounding variables like the "X-factor." In addition, a question was added that asked students to apply what they had learned previously about sampling and generalization, and consider whether results could be generalized to a broader group given how the sample was selected.

"Strength Shoe" final revisions: After this activity was modified, it was sent to the co-advisers of this project for feedback. Changes were made to improve clarity, reduce some redundancies in the software instructions, and keep terminology consistent (e.g.,

rather than using the terms “manual” and “purposeful” assignment interchangeably, the term “purposeful” assignment was kept.) The activity was then sent to the team of instructors who would be implementing it. The activity was judged to be longer than what the class would reasonably get through in one class period.

In considering what to cut, the main focus of the activity was kept at the center. This focus was to have students learn the concept that random assignment helps to balance out confounding variables on average, even though there is not necessarily perfect balance in a single random assignment. Changes were made to cut much of the reading within the activity and instead focus on the activity questions. Instead, the ideas from these omitted readings were to be addressed in the wrap-up discussion as indicated by the lesson plan. The activity draft sent to the instructors for feedback had included questions about whether students’ observed differences in average heights and in percent of subjects with the “X-factor” were unlikely to happen by chance. There were also questions asking students to determine the probability of finding a statistically significant difference in a single random assignment. These questions were taken out based on feedback from the instructional team. According to the instructors’ feedback, answering a question about determining the probability of finding a statistically significant difference would be confusing to students who had not yet learned about Type I error. It was determined that students could visualize the ability of random assignment to balance out groups by looking at the center of differences from many random assignments, without having to complicate the concept by considering the probability of a statistically significant difference.

Modifications were made to shorten the activity and make the structure and terminology consistent with what students had seen in previous class activities. While the

final “Strength Shoe” activity was still somewhat longer than a typical activity for that class, the subsequent activity (“Murderous Nurse”) was shorter than most activities, so it was determined that students would have enough time to complete both activities in two class periods. The final Strength Shoe activity for the in-class sections can be found in Appendix B3.

3.4.5 *Murderous Nurse* activity

The “Murderous Nurse” activity appeared in the original curriculum for the CATALST course (Zieffler & Catalysts for Change, 2015, p. 187-192) and was chosen for this unit because it presents a scenario that is true of many statistical studies: there is no random sampling or random assignment. However, studies like this one can still provide useful information or give a reason to investigate further. This activity happened after students had completed the “Sampling Countries” and “Strength Shoe” activities. Thus, they had been introduced to concepts of random sampling and biased vs. unbiased estimation, and of random assignment, confounding variables, and causation. Additionally, before the “Murderous Nurse” activity, students had been assigned to read the “Scope of Inferences” reading which distinguishes between random sampling and random assignment. In the “Murderous Nurse” activity, students performed a randomization test for a difference in proportions and applied what they had learned about study design and conclusions to reason about scope of inferences for this study.

The “Murderous Nurse” activity refers to the nurse Kristen Gilbert who, in the 1990s, was convicted of murdering patients. Students were presented with data from a large (non-random) sample of hospital shifts. For each shift, two variables were recorded: Whether or not Kristen Gilbert was working (yes/no) and whether or not a patient death

occurred during the shift (yes/no). Students computed the difference in percentage of deaths that occurred between the shifts Kristen was working and the shifts she was not working. Then, they were asked to think about whether this sample difference in percentages was enough evidence to say that she was killing patients, or if there could be an alternative explanation. Students used *TinkerPlots*^{TM} to set up a randomization test and found a very low p-value. Then, they were asked about whether or not results could be generalized to all shifts at the hospital, and whether or not one could conclude that Kristen Gilbert caused the deaths.

This activity did not undergo very many changes from the original, because it already addressed the desired learning goals of carrying out and interpreting a statistical test and considering the scope of inferences that could be made from the results of that analysis. The only change in content was eliminating a question that had students compare the study design in "Murderous Nurse" with study designs of other activities that they had not actually seen yet at this point in the curriculum. The "Murderous Nurse" activity was sent to the project co-advisors and to the team of instructors for feedback, but most of the changes involved minor wording edits and also the removal of detailed instructions for how to conduct the randomization test. The detailed instructions were removed because students already had experience with randomization tests for difference in proportions, so less scaffolding was needed than with previous activities. The final "Murderous Nurse" activity for the three in-class sections can be found in Appendix B5.

3.4.6 *Survey Incentives* activity

Students learning about study design can have problems distinguishing between random sampling and random assignment (Derry et al., 2000). Therefore, an activity was

necessary that would help students integrate the concepts they had learned about generalization and causation, and distinguish between the two types of randomness of random sampling from a population and random assignment to groups. As students had already seen contexts that involved random sampling only ("Sampling Countries"), random assignment only ("Strength Shoe"), and neither random sampling nor random assignment ("Murderous Nurse"), it was decided that they would next see a context in which both random sampling and random assignment could occur. A fictitious context was created based on a real study by Singer, Hoewyk and Maher (2000) in which a sample of participants was selected via random digit dialing. Some of those participants were selected to receive a monetary incentive for completing a phone survey, and the rest were in a control group that did not receive any incentive.

The fictitious setting in the "Survey Incentives" activity involved a town mayor who wanted to conduct a pilot study to determine whether a monetary incentive would increase response rates for a survey she wanted to administer about improvements that could be made to the town. Students played the role of statistical consultants in order to help the mayor design her study.

"Survey Incentives" initial draft: The "Survey Incentives" activity was drafted so that students would first go through the process of random sampling and comparing samples to the population, then go through random assignment and compare groups to each other, and finally, distinguish between the two processes.

In the first part of the activity (called "Sampling"), students first considered a biased sampling method proposed by the mayor of dropping surveys into mailboxes in her neighborhood. Students were asked to advise her on a better way to sample, given that the

mayor has a list of contact information for all town residents. Next, students used *TinkerPlots*^{TM} to take a random sample of residents, and they plotted the four variables given (sex, age, income, and hours worked per week). They took a few repeated samples and observed how they varied, and considered whether the distributions and sample statistics from each sample looked similar to the population distributions and parameters that were given to them.

In the second part of the activity (called "Assignment to Groups"), students considered confounding variables and random assignment. They were first asked to reason about what variable(s) given (sex, age, income, or hours worked per week) could be confounding variables that would affect residents' willingness to respond, and why. Then, they considered how to assign 25 participants into the two groups so that the potential confounding variables were not a concern for attempting to determine whether the incentive worked. The odd number of participants was chosen because students sometimes have the misconception that even if random assignment is used, one cannot make causal claims if the two groups are of unequal size (Wagler & Wagler, 2013). Then, similar to what they did in the "Strength Shoe" activity, students first conducted one random assignment and observed how differences in statistics vary. Next, they conducted many random assignments and plotted differences in sample statistics for one of the potential confounding variables they chose, which were centered at zero on average.

In the third part of the activity (called "Conclusions"), students considered the differences between the randomness in the "Sampling" part of the activity and the randomness in the "Assignment to Groups" part of the activity. They were told that the mayor had carried out her study using both random sampling and random assignment, and

had found that those who received the incentive were significantly more likely to respond. Then, students were asked whether the mayor could generalize her findings to the population of the town and conclude that across the town, those who receive the incentive are more likely to respond. Next, students were asked whether one could conclude that the incentive actually helped (i.e. make a causal claim). Finally, students were asked to write a short report explaining to the mayor the difference between random sampling and random assignment, and why it is not the case that as long as there is something random about the study, the mayor can make both generalizations and causal claims.

"Survey Incentives" revisions: The first draft of the activity was initially sent to the project co-advisors for feedback, and based on this feedback, one major change was made to the content in the "Sampling" part of the activity. It was pointed out that in the "Sampling" part of the activity when variables were plotted for single samples, occasionally the plots of the variables for the sample looked quite different from the plots for the population. If students found this, they might conclude that random sampling does not produce samples that are representative of the population. Therefore, a few questions were added that asked students to take many random samples and plot the sample statistics for one of the variables. This is similar to the process they followed in "Sampling Countries" by plotting the average life expectancy for many samples, except that students had the opportunity to examine the distributions of sample statistics for several different variables.

A change was made in the "Assignment to Groups" part of the activity to focus on only the three quantitative variables (age, income, and hours worked per week) as potential confounding variables to plot, rather than including sex as one of the optional variables to

plot. By focusing on quantitative variables, the activity instructions could be simplified by asking students to plot differences in means, rather than having to customize the instructions based on whether they were plotting differences in means or proportions.

Next, a revised draft of the activity was sent to the instructional team for feedback, and later back to the project's primary advisor for an additional review. The changes made based on their suggestions were mostly minor and related to making the activity more consistent in format with the activities students had previously experienced in this course. For example, longer questions were split into shorter, multiple questions, and some reading was cut. The final in-class version of the "Survey Incentives" activity can be found in Appendix B6.

3.5 Modification of activities for online class

After the in-class versions of the activities were finalized, they were modified for the online section in consultation with the online instructor of the course. For the online class, students typically completed activities with many questions, but posted responses to only a few key questions on the discussion boards. They were then required to respond meaningfully to at least one other post made by a peer or instructor, and contribute to group summaries which were constructed as GoogleDocs. The only activity in this study design unit that did not have a summary component, due to time restrictions, was "Murderous Nurse."

Given the structure of the online course, some changes were needed in the format of the activities. For each activity, key questions in the activity were identified to be "Group Discussion Questions" to which students would post responses, and in some cases new key questions were written. For the "Strength Shoe" and "Murderous Nurse" activities, the

readings required prior to these activities were placed within each activity document before the activity component. This was in order to ensure that students did the reading before the activity and to reduce the number of files that needed to be downloaded. However, the activities remained the same in learning goals and overall content.

The one activity in the study design unit that needed a significant change in structure was "Sampling Countries." This is because in the in-class version, student groups each produced a convenience sample, computed the average, and plotted it on the instructor's computer. Online, this is more difficult to do and would have required extra time for grouping together students' statistics and plotting them before they could proceed with the activity. Therefore, for the online version of "Sampling Countries," students were given 14 sample mean life expectancies for 14 hypothetical convenience samples. These hypothetical convenience samples were generated using a simulation from a *TinkerPlots*^{TM} sampler that gave countries with a higher Gross Domestic Product (GDP) a higher chance of being selected than countries with lower GDPs. Students did not see this simulation, but it was used in the development of the activity to simulate the convenience samples given. This resulted in a plot of 14 sample means centered at 72.7 years, with three sample means below the population parameter of 71 years.

Students were then asked to generate their own convenience sample of 20 countries and add their sample mean to this plot. Each student would then have a plot of 15 sample means that tended to overestimate the parameter of 71. Even if any students ended up with a sample mean lower than 71, there would still be more means above the parameter than below, and thus they would see a set of sample means produced by a biased sampling method. Students then answered key questions related to what it means for sampling to be

biased, whether random sampling appears to be unbiased, and whether it is better to take a convenience sample of size 20 than a random sample of size 10.

The activities "Strength Shoe," "Murderous Nurse," and "Survey Incentives" were nearly identical to the in-class activities, except that some questions were moved around or added to be key group discussion questions to which the students were asked to post answers on the online discussion boards. The "Strength Shoe" group discussion questions focused on comparing purposeful and random assignment, with random assignment producing, on average, groups that are equivalent with respect to known and unknown confounding variables. Students also answered wrap-up questions about whether random assignment of type of shoe would help facilitate causal claims between type of shoe and jumping ability, and about what a convenience sampling method implies for generalization to a population. For "Murderous Nurse," the group discussion questions focused on the results and conclusion of the randomization test, and the implications of the study design on conclusions that could or could not be made. In "Survey Incentives," group discussion questions focused on examining plots of statistics for different variables obtained via random sampling, examining plots of differences in statistics for different variables when random assignment was used, and contrasting the implications of random sampling versus random assignment for making conclusions. The online class versions of the "Sampling Countries," "Strength Shoe," "Murderous Nurse," and "Survey Incentives" activity can be found in Appendix C.

3.6 Development of the IDEA assessment

In order to measure student learning outcomes before and after the study design curriculum, the IDEA was developed to use as a pretest and posttest. First, a test blueprint

was developed containing sixteen different learning goals (Appendix I). Eight of these learning goals are related to sampling and generalization, and the other eight are related to assignment to groups and establishing causation. These learning goals were chosen based on concepts that were emphasized in many of the statistics textbooks reviewed (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009; Devore & Peck, 2005; Lock et al., 2012; Moore, 2001, 2010; Moore & McCabe, 1999) and on the limited statistics education research that exists about students' understanding of generalization and causation (e.g., Derry et al., 2000; Wagler & Wagler, 2013).

After the development of the blueprint, many existing assessments of introductory statistics that already have some evidence of content validity were examined in order to find items that would align, or could be easily modified to align, with the identified learning goals. The existing assessments reviewed include the Comprehensive Assessment of Outcomes in Statistics (CAOS; delMas, Garfield, Ooms & Chance, 2007), Goals and Outcomes Associated with Learning Statistics (GOALS; Sabbag, 2013; Sabbag & Zieffler, 2015), Basic Literacy in Statistics (BLIS; Ziegler, 2014), and Levels of Conceptual Understanding in Statistics (LOCUS; http://education.ufl.edu/locus/). In addition, items related to study design from the Assessment Resource Tools for Improving Statistical Thinking (ARTIST; Garfield et al., 2002) website were reviewed.

For all but three learning goals, an item or set of items was found in an existing assessment that addressed the learning goal, or could be slightly modified to address it. For two of these three learning goals, open-ended items were found in the ARTIST website which could be modified into forced-choice items to address the learning goal. The only learning goal for which no existing items were found was "Ability to distinguish between

statements that make causal claims and statements that make association-only claims." For this learning goal, a set of items was created using existing media article headlines. Students had to determine whether each headline was making a statement of association only or a statement of causation. Table 3.3 shows the learning goals and sources of the items that were modified from existing items on various assessments. Almost all items were modified at least slightly from their original version, but some new items were drafted based on existing item contexts.

Table 3.3
Learning outcome and original source of each item on the IDEA Assessment

Item	Learning Outcome	Source
1-2 (item set)	Ability to identify the sample and the population to which inferences can be made	BLIS – Item 1
3	Ability to understand what it means to make an appropriate generalization to a population, using sample data	ARTIST Item Database – Item Q2027[a]
4	Ability to understand the factors that allow (or do not allow) a sample of data to be representative of the population	CAOS Item 38/BLIS Item 35
5	Ability to understand when sample estimates may be biased due to lack of a representative sample	GOALS v.2 – Item 2[b]
6	Ability to understand that a small random sample is preferable to a larger, biased sample	ARTIST Item Database – Item Q2442[a]
7	Ability to understand that random sampling is preferable to non-random methods of sampling for a sample to be representative of the population	LOCUS item
8	Ability to understand that sample statistics vary from sample to sample	BLIS - Item 8
9	Ability to recognize that random sampling is the most salient issue when using a sample to generalize to a population	ARTIST Item Database – Item Q1237[a]
10	Ability to determine what type of study was conducted (observational or experimental)	ARTIST Topic Scale Test – Data Collection – Item 7
11	Ability to understand that a randomized experiment is needed to answer research questions about causation.	ARTIST Topic Scale Test – Data Collection – Item 4

Item	Learning Outcome	Source
16	Ability to understand that correlation does not imply causation.	CAOS item 22/ GOALS v.2 – Item 3[b]
17	Ability to understand how a confounding variable may explain the association between an explanatory and response variable	ARTIST Topic Scale Test – Data Collection – Item 9
18	Ability to understand the purpose of random assignment in an experiment: To make groups comparable with respect to all other confounding variables.	CAOS - Item 7/GOALS v.2 – Item 1
19-21 (item set)	Ability to understand that random assignment is the best way to balance out groups with respect to confounding variables.	Item #5 in ARTIST Topic Scale Test – Data Collection, modified by Wagler & Wagler (2013)
22	Ability to recognize when a randomized experiment is the most salient research design for a particular research question.	LOCUS item

Note. All items were modified from existing constructed response items except for the following:

a. Forced-choice item drafted based on context from original free-response item.
b. Original item used from GOALS v. 2 (no modification).

In modifying and creating the IDEA items, item writing guidelines provided by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and Haladyna, Downing, and Rodriguez (2002) were considered. For example, the central ideas were included in the stem rather than the choices, and care was taken to make sure that the correct answer was not the longest answer.

After an initial draft of the assessment was created, it went through a first round of feedback by the project co-advisors, who are highly experienced in assessment development, especially as relates to statistics education research. Modifications were made for clarity, and some multiple-choice items were split into smaller item sets in order to ease cognitive load and also to test for multiple possible misconceptions. For example, one item originally asked students to select the response option that represented an appropriate way to randomly assign participants into four groups. Rather than having this appear as a single item where students selected one of three options, the item was instead

split into three smaller items (see items #19-21 in IDEA, Appendix J). For each item, students selected whether or not the described method of assigning participants to groups was appropriate for balancing out confounding variables.

After modifications were made based on this first round of feedback, three external statistics education experts were invited to review the blueprint and assessment. All three reviewers had an extensive amount of experience teaching statistics and were well known in the field of statistics education research. Two of the reviewers were authors of introductory statistics textbooks. An invitation e-mail was first sent to each of these three statistics education experts to explain the purpose of the study and ask if they would be willing to give feedback on the blueprint and assessment that would be used as a pretest and posttest (Appendix H1). All three experts agreed to review the blueprint and assessment.

The reviewers were then sent a copy of the blueprint and assessment. In the instructions (Appendix H2), they were asked to give any suggestions they had for improving an item, keeping in mind the intended learning goal in the assessment blueprint. They were also asked to give feedback on clarity and wording of the items and response options. Additionally, the expert reviewers were asked to review the blueprint and give feedback on whether any learning goals were missing or redundant.

Only one of the three reviewers gave suggestions on the blueprint. These dealt with clarity and wording issues, and only two minor wording changes were made to two of the learning goals based on this feedback. The final blueprint can be found in Appendix I. The feedback on the assessment from the three reviewers consisted mainly of suggestions for wording to make items clearer. None of the reviewers indicated that any items were

misaligned with their corresponding learning goals, or that any learning goals were missing or redundant.

The IDEA assessment was revised based on feedback from the three reviewers and an additional round of feedback from the main project advisor. The final assessment had 22 total items measuring 16 learning goals. Nine of these items, covering 8 learning goals, were related to sampling and generalization, and will be referred to as the "sampling" items. The other 13 items covered the remaining 8 learning goals were related to random assignment and causation, and will be referred to as the "assignment" items. The final version of IDEA appears in Appendix J.

3.7 Development of group quiz and lab assignment

Two assignments consisting of constructed-response questions were created to assess students' understanding of study design and conclusions. In the CATALST course, students were randomly assigned to groups, and they would work on activities and take quizzes together. For the in-class sections, groups consisted of two to three students each. For the online section, groups consisted of four to six students and were maintained the same throughout the semester. Groups were rotated periodically in the in-class sections, but students in the same group took quizzes together after already having worked together on at least one or two activities. For the in-class sections, a group quiz was given on the class period following the "Murderous Nurse" activity for the in-class sections. For the online section, the group quiz was due following the "Survey Incentives" activity. A homework assignment (referred to as a "lab assignment" in this course) was also created to be completed individually after students had completed all of the activities in this study design curriculum.

3.7.1 Group Quiz

The group quiz was created to assess students' ability to transfer their conceptual knowledge about study design and conclusions to headlines and claims made from studies. Three contexts were used: a *Gallup* media article describing a real study linking moderate drinking to improvement in mental health (Nekvasil & Liu, 2016), a real study linking larger bowls to people's tendency to serve themselves larger quantities of ice cream (Wansink, Van Ittersum, & Painter, 2006), and a hypothetical study finding a link between higher GPAs and higher chances of getting into U.S. medical schools. The studies in the first and third scenarios involved random sampling but not random assignment, and the study in the second scenario involved random assignment, but not random sampling. Each context had two questions, one relating to generalization and one relating to causation.

Modifications were made to the quiz based on feedback from the project advisors and instructional team, but most changes were minor. The teaching team indicated that they typically designed their group quizzes to stretch and challenge students, so some questions were changed to give less scaffolding. For example, rather than asking students why a given headline was appropriate given that random sampling was used in the study, a question was rephrased to simply ask if the headline was appropriate given the study design, and why. Additionally, based on feedback from both the advisors and the instructional team, some headlines were rephrased so that they could be more clearly identified as making either a generalization or a causal claim. Some of the reading from the stems of the questions was cut to reduce length and present only the necessary information. The final Group Quiz 5 appears in Appendix F1 and was the same for both the in-class and online sections.

3.7.2 Lab Assignment

The lab assignment was modified from an existing lab assignment in the course, which presented summaries of three different studies relating to peanut allergies in children. For each study, students had to identify the treatment and response variables, describe what a "significant result" means, consider how the sample was selected and implications for conclusions, and consider how groups were assigned (or not assigned) and the implications of the assignment method for making conclusions.

This existing lab assignment was modified to present two studies instead of three so that more questions could be asked about study design and conclusions for each study. Rather than using the term "treatment variable(s)" as the original lab assignment did, the revised lab asked students to identify the "explanatory variable(s)" (as well as the response variable) because not all of the studies involved were experiments. Students were asked to indicate whether the study design allows researchers to generalize to a wider population. Students were still asked to consider what statistical significance meant, as this was an important idea of the course even though it was not a central topic in the study design unit. Next, students were asked to consider whether participants had been assigned (or not assigned) to groups and what that implied for potential conclusions.

One question was added to test for potential confusion between random sampling and random assignment. This question had students explain whether or not it was appropriate to conclude that random sampling of pregnant women would allow for causal conclusions to be made between peanut consumption during pregnancy and peanut allergies, thus testing for possible confusion between random sampling and random assignment. Another question was added to promote students' critiquing of the study

design of an observational study with no random assignment. This very last question asked students to reason about whether the study described would provide justification for someone to avoid eating peanuts during pregnancy to prevent the infant's peanut allergies.

After the initial draft of the group quiz and lab assignment were each formed, they were sent to the advisors and the instructional team for feedback. Some questions were added for scaffolding and for clarity. For example, instead of asking students whether they could generalize to a wider population, a more specific question was added asking them to first identify the population of interest. Then, students were asked whether or not they could generalize to that population, given the study design. Also, some question wording was changed to be clearer and easier to understand, and terminology was changed to be consistent. For example, rather than asking about how participants were assigned or not assigned into "groups/conditions", the term was changed to just "groups" because some of the explanatory variables were not controlled. The final lab assignment is found in Appendix G1 and was identical for both the in-class and online sections.

3.7.3 Rubrics

After the group quiz and lab assignment were finalized, rubrics were created for each assignment. Group quizzes in this course typically had detailed rubrics, with each question worth a certain number of points and partial credit available. Therefore, for this group quiz, a rubric was written indicating components required to earn a full point for each question. The quiz was scored out of 6 points, one point per question. Multiple scenarios for earning a half point were included for each question. For example, if students misinterpreted a claim made in a headline but reasoned correctly about the study design and scope of inferences given their interpretation, they would get half a point for that question. The

rubric was reviewed during a meeting with the instructional team, and some changes and additions were made to be more specific about components needed and allow for more potential partial credit scenarios. The group quiz rubric appears in Appendix F2. After the rubric was finalized, it was given to the two teaching assistants who graded the in-class quizzes. For the online class, the instructor graded the quizzes.

The lab assignments in the CATALST course were typically graded holistically according to the following scale:

> (3) Answers exhibit a **complete understanding** of the concepts in the assignment. There are no errors in student's statistical reasoning. The responses are clear and correct.
> (2) Answers exhibit a **near complete understanding** of the assignment. There are perhaps minor errors in student's statistical reasoning or the responses are slightly unclear or incorrect.
> (1) Answers exhibit **some understanding** of the assignment. There are errors in student's statistical reasoning or the responses are unclear or incorrect.
> (0) Answers exhibit **little to no understanding** of the assignment. There are fundamental errors in student's statistical reasoning or the responses are unclear or incorrect.

The instructors requested that a rubric be written for the lab assignment containing a bullet list of the main points that students were supposed to understand. A rubric was created with a list of concepts that students should understand, which addressed most of the questions in the lab. However, the instructors then suggested that the list be narrowed down to at most three to four main concepts. The concepts were then narrowed down to the most important ones, focusing on study design and conclusions. Therefore, there were some questions on the lab that were now deemed less important and would not be considered for the holistic grade. For example, students were asked what it meant for

results to be "statistically significant," but as this was not directly relevant to study design, it was not included as one of the major concepts to consider for holistic grading. The final lab rubric (found in Appendix G2) was given to the instructors and also shared with the teaching assistant for the online class, who graded the online labs. The two in-class instructors graded their students' lab assignments.

3.8 Implementation of unit

The study design unit, consisting of four activities, a group quiz, a lab assignment, and the IDEA pretest and posttest was implemented in the CATALST course, an undergraduate introductory statistics course at the University of Minnesota. The unit lasted two and a half weeks in the second half of spring semester 2016. Lesson plans were developed, and instructors were trained on each activity at least several days prior to the class period for that activity.

3.8.1 Lesson plans

One lesson plan was developed for each of the four activities: "Sampling Countries," "Strength Shoe," "Murderous Nurse," and "Survey Incentives." The lesson plans each began with a summary of the activity, learning goals, preparation requirements for students (e.g., assigned readings), and names of the *TinkerPlots*^{TM} files needed. Teacher instructions were also provided for each part of the activity, along with suggestions for approximate time periods to spend on each portion of the activity. The lesson plan contained questions for instructors to ask during large group discussion times. The most important questions were highlighted in case time did not permit for discussion of all questions. Also, the lesson plans contained suggestions for potential issues that the researcher anticipated for the students, such as questions that were thought to be more

difficult for students. These suggestions contained potential questions to ask to give students more scaffolding. While the lesson plans were written to target an in-class environment, the online instructor also received the lesson plans in order to be aware of the main learning goals and key discussion questions as he structured his activity wrap-ups. All lesson plans can be found in Appendix D.

The researcher met with the team of instructors at various times during their regular hour-long weekly meetings. First, near the beginning of the semester, there was a meeting to explain what the curriculum would entail, and a schedule was made for developing the activities, giving feedback, and finalizing the activities. Two weeks prior to the start of the unit, the researcher began attending the weekly meetings. At the first of these meetings, feedback was given on the first activity and the structure of the lesson plans was discussed. At subsequent meetings, the researcher went over the activities and lesson plans (which the instructors had already received at least several days prior to the meeting). Instructors had the opportunity to ask questions about the activities and lesson plan components or give additional feedback on the lesson, which often resulted in minor edits to the lesson plan. During some of the meetings, instructors also discussed the group quiz and lab rubrics, and suggested edits and additions to these.

All meetings were attended by the researcher, the two in-class instructors, and occasionally by the coordinator of the course who was also a co-advisor on this project. Due to schedule conflicts, the instructor for the online class often arrived during the second half of the meetings, but he also consulted with the researcher during the time that they met to revise the activities for an online format. The teaching assistants typically did not attend

these meetings except for the teaching assistant for section 2 who attended the meeting where “Sampling Countries” was presented.

At the last two meetings for this unit, there was also some debriefing where instructors reported their impressions on how the activities went, what progress was seen in students’ understanding, and what could have gone better. At the debriefing, the group discussed potential changes to the course curriculum and placements of the study design topics, as well as concepts and tasks that students appeared to struggle with on their quiz and lab assignment.

3.8.2 Class observations

The three in-class sections were observed by the researcher and a co-observer who was another graduate student in statistics education. A total of fifteen 75-minute classes were observed (five class days for each of the three sections). The class was videotaped, with the camera pointed at the instructor. The video focused on the large group discussions, and even though the camera was kept rolling during activity time, students’ small group discussions were not audible due to the large number of groups. Instead, observers circulated around the room while student groups worked on the activities. During this time, observers took notes on discussions that they heard among student groups and on interactions between students and instructor. In general, the researcher and co-observer tried to be in different areas around the room so that they could hear different groups’ discussions.

The researcher and co-observer used an observation form for each of the four activities (see Appendix E), checking off the large group discussion questions on the lesson plan if the instructor asked them. Also, they checked off any of the anticipated potential

issues on the checklist for students, and potential suggestions that the instructor used to deal with these issues that they observed. The researcher and co-observer also visited the classroom during the group quiz, but no observation form was made for the quiz day because no lesson plan was necessary for the quiz. Instead, the observers focused on listening to group conversations. They took notes on how students appeared to be reasoning for each of the questions, what they seemed to be understanding, and where they seemed to be struggling.

The online class had no face-to-face meetings, so the observation for this class consisted of reading the discussion boards where students posted answers to questions. The researcher also read student groups' summaries for each of the three activities that had a required summary: "Sampling Countries," "Strength Shoe," and "Survey Incentives." In addition, the researcher read the wrap-up announcements posted by the online instructor and also watched a video that the instructor made to wrap up "Sampling Countries."

After the conclusion of the study design unit, the researcher watched the videos of the in-class sections. Only large-group, and not small-group, discussions were audible on the videos, but observers had documented their observations of small-group discussions. Notes from the observations and videos were typed and summarized. The findings from all class observations are summarized in the Results chapter.

3.8.3 Group quiz administration

In between the "Murderous Nurse" and "Survey Incentives" activities, the in-class sections took a group quiz written for this unit (see Appendix F1 for the quiz). The researcher and co-observer visited the class on the day of the quiz and took notes on student discussions. Because there was no lesson plan for the day of the quiz, there were no

observation checklists. The researcher and co-observer simply took notes on what they observed students discussing correctly, and where they seemed to be having problems.

There were some differences in how the quiz was administered between the in-class sections and the online section. In class, students took the quiz in the groups of 2-3 students with whom they had been working for previous activities in the unit. There were a total of six questions on the quiz, and each pair of questions was required to have a student who was the writer or recorder of the answers. Online, students took the quiz on a GoogleDoc in the group of 4-6 students with whom they had been working on online discussions. Since the groups were larger, each individual question was required to have one main author who composed the answer on the GoogleDoc. Students then edited responses when necessary and also had the capability of using the "Comments" feature to discuss answers. However, only one of the eight total groups in the online section used the "Comments" feature to discuss answers. While one additional group used comments, the group members used them only to ask members to indicate if they disagreed with a particular answer, and nobody indicated disagreement in the comments. It is unknown whether any of the groups discussed answers over e-mail or chat forums.

In sections 1 and 3, the instructor handed out the quiz without any prior discussion. However, the instructor of section 2, who had run out of time for wrap-up on the "Murderous Nurse" activity, decided to lead a brief large-group discussion at the beginning of the group quiz period.

3.9 Data analysis

After the implementation of the study design unit, data from the students' assessments (IDEA pretest, IDEA posttest, group quiz and lab) were analyzed. First, in

order to de-identify the data, a unique number was assigned to each student. Numbers were assigned so that the first digit of each student's ID would indicate their section number, the second digit (or pair of digits) would indicate a randomly assigned group number, and the last digit would make the number unique to each student. For example, students in the 10th randomly ordered group of section 1 had IDs 1101, 1102, and 1103, and students in the 7th randomly ordered group of section 4 (the online section) had IDs 471, 472, 473, 474, and 475.

As the completion of the IDEA pretest and IDEA posttest was included as part of students' grades, almost all students completed them. Two students overall (one in section 1 and one in section 3) returned a signed consent form indicating that they declined to participate in the research, so their responses were deleted. One student from section 2 was a retired statistics instructor auditing the class, and this student's IDEA and lab responses were also deleted because he did not represent the target population of undergraduate introductory statistics students. All but two students completed 20 or more of the 22 items. These two students had completed only two items (one student on the pretest and the other student on the posttest), and thus these two cases were also deleted from the dataset. There were several duplicate attempts for either the pretest or posttest. If one attempt was partial and another complete, the partial attempt was deleted. In the cases where duplicate attempts were both complete, the one that took a longer amount of time was kept, and the short attempt was deleted.

After this data cleaning was done, there were 140 total students whose data was eligible for analysis, split among the four sections as shown in Table 3.4 below. This table also shows the percentages of students in each section who completed each assessment,

and the percentage of students for whom the IDEA pretest, IDEA posttest, quiz, and lab are all available for analysis. Sections 1 and 3 have a lower percentage of students with complete data than the other sections, but this is partly because group quizzes were omitted from the analysis if one group member did not consent to be in the study. One student in section 1 and two students in section 3 did not have their quizzes analyzed because they worked with non-consenting group members. If these students had not had their quizzes omitted from analysis, then there would be 85.7% of students in section 1 with complete data and 79.3% of students in section 3 with complete data.

Table 3.4
Percent of total eligible students [a] *completing unit assessments for four sections of EPSY 3264*

	Percent of Section				Percent of
	1	2	3	4	Total
	n = 42	*n* = 33	*n* = 29	*n* = 36	*n* = 140
Completed IDEA pretest	92.9	97.0	82.8	100	93.6
Completed IDEA posttest	92.9	97.0	96.6	91.7	92.9
Completed both IDEA pretest and posttest	90.5	90.9	82.8	91.7	89.3
Completed group quiz	100[b]	100	100[b]	94.4	98.6
Completed lab assignment	92.9	93.9	96.6	91.7	93.6
All assessments available for analysis[c]	83.3	87.9	72.4	88.9	83.6

[a] Eligible students represent all students who consented to participate in the study and were considered to be representative of the target population of introductory statistics students.
[b] Although 100 percent of students completed the group quiz in section 1 and 3, one student in section 1 and two students in section 2 were in a group with a class member who declined to participate in the study, so those students' group quizzes were excluded from the analysis.
[c] The last row of the table indicates the percent of eligible students who have a complete set of assessments available for analysis. This excludes students whose group quiz was taken out of the analysis due to lack of consent of a group member.

3.9.1 Quantitative data analysis

In order to answer the research question (see section 3.1), scores on the IDEA assessment were compared from pretest to posttest. Total scores on the assessment (number of items correct) were examined. In addition, sampling subscores (number of sampling items correct) and assignment subscores (number of assignment items correct) were examined.

In order to measure reliability of the assessment, a measurement expert at the University of Minnesota was consulted in order to ensure appropriate analyses were conducted. Omega coefficients (MacDonald, 1999) were computed in order to measure reliability for the total score and for each subscore, for both pretest and posttest. The *psych* package in R was used to compute the omega coefficients (Revelle, 2017).

Descriptive statistics were computed for the total score and for each of the sampling and assignment subscores. Paired *t*-tests were conducted to test whether changes from pretest to posttest were statistically significant for the total score and for each subscore. In addition to examining scores of the full sample of students, the scores were also examined separately for each section. One-way ANOVA tests using multiple comparison adjustments were conducted to examine whether there were any significant differences between each pair of sections on pretest scores, posttest scores, and changes in score from pretest to posttest.

Additionally, the response distributions for each item were compared from pretest to posttest. McNemar's test with multiple comparison adjustments was used to determine whether or not the change in percent of correct responses for each item was statistically significant. Changes from pretest to posttest were also examined for item sets.

3.9.2 Development of codes used for qualitative data analysis

Responses to the group quiz and lab assignment were analyzed qualitatively. In order to do this, a coding system was created in order to categorize students' responses. Of interest in this project was examining areas in which students appeared to display correct understanding of ideas in study design and conclusion, and areas in which they still held misunderstandings or misconceptions after experiencing activities in the curriculum. As student papers were examined, it was discovered that some contained responses that made it ambiguous to the reader whether or not students displayed a correct understanding of the distinctions between the purposes of random sampling and random assignment. For example, some students wrote answers stating that both random sampling and random assignment are needed to make both generalizations and causal claims, when only asked about one conclusion or the other. This type of answer did not make it clear whether students had a correct understanding of the differences between generalizing to a population and making causal claims. Therefore, three categories of behaviors to code were developed: (1) Incorrect understanding, (2) Correct understanding, and (3) Ambiguity. The "Ambiguity" category refers to behaviors in which it is not clear to a scorer whether a student has correct understanding about a concept. The full codebook, along with student examples, can be found in Appendix L.

Since the lab assignment involved a single context, the lab assignment answers were coded as a whole (with the exception of some codes specific to the last two questions, discussed below). The group quiz presented students with three different contexts. For each context, there was one question related to generalization to a population and one related to making causal claims. The three scenarios were each coded separately.

Codes were developed for each of the three categories described above, and each code was given a label to facilitate reference to the code. Each code label begins with a letter corresponding to its category (I = Incorrect, C = Correct, A = Ambiguity). Some codes, in particular those related to misconceptions, were developed based on what research literature has said about students' understanding of study design and conclusions. Other codes, in particular those related to correct understanding, were derived from guidelines of what students should understand regarding study design and conclusion based on reviews of textbooks and other documents (e.g., GAISE, 2016) and reviews of introductory statistics textbooks. Some codes were inspired from behaviors observed during the classroom observations of the activities, while others emerged while reading answers during the coding process. If a code emerged during the coding process, all assignments that had previously been read were re-read to see if the new code was present. Table 3.5 below summarizes the codes created and the sources that inspired the development of each code.

Table 3.5
Behaviors used for qualitative analysis coding, along with labels and sources that inspired the development of each code.

Code label	Coded behavior	Source
[I]	*Misconceptions/Incorrect Thinking*	
[I-TC]	*Misunderstandings about which study designs help with which types of conclusions*	
I-TC-RSC	Bringing up only random sampling/lack thereof when the question is about causation	Derry et al. (2000); delMas et al. (2007); Tintle et al.. (2012); Sabbag (2013); Classroom observations
I-TC-RAG	Bringing up only random assignment/lack thereof when the question is about generalization	
I-TC-BOTHG	Saying you need both random sampling AND random assignment to generalize	
I-TC-BOTHC	Saying you need both random sampling AND random assignment to make causal claims	
I-TC-CLAIM	Confusing the meaning of "generalize" with the meaning of "causal claims"	Classroom observations

Code label	Coded behavior	Source
I-TC-NOCC	Not believing causal claims can be made even though random assignment was used	Sawilowsky (2004)
[I-SS]	*Incorrect beliefs about sample size*	
I-SS-UNEVEN	Saying that unequal sample sizes in two groups do not allow for any conclusions	Wagler & Wagler (2013)
I-SS-LARGEN	Saying we can generalize due to the large sample size	delMas et al. (2007); Derry et al. (2000)
I-SS-SMALLN	Saying we can't generalize (or make any conclusion) only because of small sample size	Sabbag, (2013); Wagler & Wagler (2013)
[I-SD]	*Difficulty understanding study descriptions*	
I-SD-RECRS	Difficulty recognizing from study description whether random sampling was used	Classroom observations
I-SD-RECRA	Difficulty recognizing from study description whether random assignment was used	
[C]	*Correct Thinking*	
[C-SG]	*Makes connections between sampling and generalization*	
C-SG-RSGEN	Pointing out that random sampling is relevant for generalizing to a wider population	Guidelines/desired learning outcomes; Classroom observations
C-SG-SCHAR	Mentioning that the sample can have characteristics that make it different from the population (if no RS was used)	
[C-AC]	*Makes connections between random assignment and causation.*	
C-AC-RACC	Pointing out that random assignment is relevant for making causal claims	Guidelines/desired learning outcomes; Classroom observations
C-AC-CONFV	Mentioning that confounding variables can make two groups different from each other (if no RA was used)	
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions*	
C-WHY-RS	Explaining why random sampling helps us to generalize	Guidelines/desired learning outcomes
C-WHY-RA	Explaining why random assignment helps us to make causal claims	
[C-EXT]	*Correct answers, but bringing in extraneous information*	
C-EXT-RS	Bringing up issues of generalization and/or random sampling extraneously when the question is about causation, while still correctly addressing the need for random assignment to make causal claims.	Classroom observations
C-EXT-RA	Bringing up issues of causation and/or random assignment extraneously when the question is about generalization, while still correctly addressing the need for random sampling to make generalizations.	
[A]	*Ambiguity (Scorer may have difficulty judging whether or not student has a correct understanding.)*	

Code label	Coded behavior	Source
A-BOTH	Does not separate generalization and causation, saying you need both random sampling and random assignment to conclude generalization and causation.	Classroom observations
A-RAND	Being vague about what kind of randomness is needed to generalize or make causal claims	Kaplan, Rogness, & Fisher (2014); Classroom Observations
A-RSNORA	Saying that only random sampling was used, thus implying that random assignment was not used	Emerged during coding
A-RANORS	Saying that only random assignment was used, thus implying that random sampling was not used	Emerged during coding

Development of codes for incorrect thinking: As seen in Table 3.5 above, 11 codes were developed to represent potential misunderstandings or incorrect thinking. These were divided into three categories: Misundertandings about which study designs help which types of conclusions (*I-TC*), incorrect beliefs about sample size (*I-SS*), and difficulty understanding study descriptions (*I-SD*).

The first category involved misunderstandings about the purpose of each study design, and what conclusions can be made. Previous research (e.g., Derry et al., 2000) and results on assessments taken by introductory statistics students in different populations (e.g., delMas et al., 2007; Tintle et al., 2012) have revealed that students tend to confuse the different purposes of random sampling and random assignment for making conclusions about statistical studies. Also, results from administrations of the CAOS test (delMas et al., 2007; Tintle et al., 2012) and the GOALS test (Sabbag, 2013) showed similar confusion, such as many students answering that the purpose of random assignment is to make the sample representative of the population. This led to the development of codes I-TC-RSC and I-TC-RAG, which would each represent matching the wrong study design with the

wrong type of conclusion. (These types of behaviors were also observed during some class activities, according to observation notes.)

In addition, it was anticipated that confusion between random sampling and random assignment could also lead to students claiming that both random sampling and random assignment are needed only for making generalizations, or only for making causal claims. This led to the development of codes I-TC-BOTHG and I-TC-BOTHC. According to classroom observation notes, some students confused the meanings of the word "generalize" and the phrase "make causal claims," which inspired the addition of code I-TC-CLAIM. It was anticipated that some students might also have disbelief in the ability of random assignment to, on average, balance out confounding variables, as documented by Sawilowsky (2004). This led to the I-TC-NOCC code to represent students not believing causal claims can be made even while acknowledging that random assignment was used.

Another category of incorrect thinking was over-emphasis of sample size rather than method of study design. Based on results found in the study by Wagler and Wagler (2013), it was anticipated that students might say that conclusions (whether these be generalizations or causal claims) cannot be made from studies where two groups have unequal sample sizes (code I-SS-UNEVEN). Another common misconception found in previous studies and assessment data is that large sample sizes allow for generalization to a population, and small sample sizes do not. For example, Derry et al. (2000) reported that students were often more concerned about issues of sample size than issues of sampling method. In assessment data from CAOS and the similar GOALS test, many students indicated that a random sample of size 500 was inappropriate for representing a population of 5,000, despite the fact that random sampling was used (delMas et al., 2007; Sabbag,

2013). This prompted the development of codes I-SS-LARGEN and I-SS-SMALLN, each representing the over-emphasis of large sample sizes, and small sample sizes, respectively.

One category of misunderstanding was students not being able to recognize when random sampling or random assignment were used, according to data descriptions. In particular, this problem was documented in classroom observation notes during the "Murderous Nurse" activity. For example, the description of "Murderous Nurse" data does not explicitly mention that the shifts are not randomly sampled nor randomly assigned, but some students had difficulty recognizing this information. It was thus predicted that students might also have difficulty recognizing from data collection descriptions whether random sampling and/or random assignment were used, leading to the codes I-SD-RECRS and I-SD-RECRA.

Development of codes for correct thinking: Eight codes were developed to represent correct thinking. Four categories of correct thinking arose: Understanding that random sampling is relevant for generalization *(C-SG)*, understanding that random assignment is relevant for making causal claims *(C-AC)*, including answers with more depth, such as why each study design leads to given conclusions *(C-WHY)*, and correct answers that bring in extraneous information *(C-EXT)*.

When developing codes that would represent correct thinking, guidelines for statistics education (e.g., GAISE, 2016) and textbooks (e.g., Agresti & Franklin, 2009; DeVeaux, Velleman, & Bock, 2009; Moore, 2010; Lock et al., 2013) were reviewed for content that many statistics educators agree students should know about study design and conclusions. For example, two important content areas that appeared in the guidelines and textbooks were the relevance of random sampling for obtaining a representative sample

that could be used to make inferences about a population, and the relevance of random assignment for balancing out confounding variables and helping to support causal claims. Therefore, the codes C-SG-RSGEN and C-AC-RACC were developed to represent a correct understanding of random sampling being relevant to generalization, and random assignment being relevant to causation. According to classroom observation notes, sometimes students correctly discussed the need for random assignment for making causal claims, when they were looking at a part of the activity that discussed generalization to a population. Thus, students sometimes brought in extraneous, albeit correct, information about a study design that was not relevant to the question at hand. This led to the development of codes C-EXT-RS and C-EXT-RA to represent this behavior of adding extraneous information, even when not directly being asked about it.

Some textbooks and activities went into more depth than others, allowing students to visualize how random sampling tended to produce unbiased estimates and how random assignment tended to balance out differences between groups, on average (e.g., Rossman et al., 2007; Zieffler et al., 2015). Similarly, some notes taken during classroom observations explained that occasionally, when students were asked whether a given conclusion could be made and why, instead of just mentioning the study design needed, they would go into depth about why the study design helped with that type of conclusion. For example, students in the classroom sometimes talked about how random assignment tended to balance out differences between the groups, so that the only real difference was the explanatory variable. Answers that explained this link between random assignment and the ability to make causal claims are arguably richer than answers that merely said that random assignment helped to enable causal claims. Similarly, answers that explained that

random sampling tends to produce samples that are representative of the population and tend to provide unbiased estimates are arguably richer than answers that merely stated that random sampling helped to enable generalizations. Therefore, codes C-WHY-RS and C-WHY-RA were developed to represent behaviors in which students would explain why a given study design was linked to a given type of conclusion.

Many of the textbooks reviewed (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009; Devore & Peck, 2005; Moore, 2010; Lock et al., 2013) discussed how sampling bias hinders generalization and how confounding hinders causal claims. The group quiz and lab assignment were designed with the recognition that sometimes there were multiple ways of correctly explaining why a certain type of conclusion could *not* be made. For example, when asked whether generalizations could be made about a study without random sampling, a student could answer correctly by either pointing out the lack of random sampling, and/or discussing how the sample was not representative of the population. This led to the development of code C-SG-SCHAR, which represented a correct answer about why generalizations cannot be made based on characteristics that make the sample different from the population. Also, when asked whether causal claims could be made from an observational study, a student could answer correctly by either pointing out the lack of random assignment, or by mentioning how other differences between the two groups (confounding variables) could explain any associations found. This led to the development of code C-AC-CONFV, which represented a correct answer about why causal claims cannot be made, based on confounding variables that could otherwise explain associations found.

Development of codes for ambiguity: At times, it was ambiguous whether students had a correct understanding of the roles of random sampling and random assignment in making conclusions. Therefore, a coding category was created for behaviors that indicated this ambiguity (A). Previous research has found student misuse of the word "random," as well as difficulties understanding the long-term behavior of randomness (Kaplan, Fisher, & Rogness, 2009; Kaplan, Rogness, & Fisher, 2014). Additionally, during classroom observations, sometimes students were observed saying that a study was "not random" or "not randomized." This made it difficult for observers to know whether students were referring to the correct type of randomness (either random sampling or random assignment) for making a given type of conclusion. This led to the code A-RAND, which represented the behavior of vaguely referring to randomness without being specific about the type of randomness.

According to observer notes during the "Survey Incentives" activity, some students stated that with random sampling and random assignment, generalizations and causal claims could be made. A statement like this, while correct, did not make it entirely clear whether students were recognizing the purpose of random sampling (enabling generalizations) as distinct from the purpose of random assignment (enabling causal claims. This led to development of code A-BOTH.

At the beginning of the coding process, student answers emerged stating that "only" random assignment was used, thus implying that random sampling was not done and thus generalizations could not be made. Similarly, other student answers stated that "only" random sampling was used, thus implying that random assignment was not done and thus causal claims could not be made. Answers like these, while not incorrect, made it

ambiguous whether students understood that it was possible to have both random sampling and random assignment in a study, or whether students incorrectly believed that one design could not happen without the other. This led to the development of codes A-RSNORA and A-RANORS.

Development of codes specific to the lab assignment: Although the lab assignment was graded holistically and used a single context throughout (consumption of peanuts and peanut allergies in infants), two questions at the end were of interest for close examination. Codes were created to examine students' behavior answering these questions, as outlined in Table 3.6 below.

Table 3.6
Behaviors used for qualitative analysis coding, specific to lab assignment

Code label	Coded behavior	Source
Question 13: Whether random sampling would allow for causal claims		
I-LAB13-RSCC	Incorrectly agreeing that random sampling allows for causal claims	
C-LAB13-RSGEN	Correctly mentioning that random sampling only helps with generalization	Derry et al. (2000)
C-LAB13-RACC	Correctly mentioning that random assignment would be needed for making causal claims	
Question 14: Making conclusions based on study with no random sampling or assignment		
C-LAB14-NOCC	Mentioning the lack of ability to make causal claims (or pointing out that random assignment was not used, or that confounding variables could explain peanut sensitivity)	Derry et al. (2000); Groth (2006)
C-LAB14-NOGEN	Mentioning the lack of ability to make generalizations (or pointing out that random sampling was not used, or that the sample may not be representative of the population)	Derry et al. (2000)
I-LAB14-PVAL	Makes a decision based only on the *p*-value, without consideration of study design	delMas et al. (2007); Sabbag (2013)
I-LAB14-NOSD	Makes a decision based on factors not related to study design (e.g., on prior contextual knowledge)	Derry et al. (2000) Wroughton et al. (2013)

Lab Question #13 presented a hypothetical student who claimed that if the study had used random sampling, this would enable causal claims. Students were asked to indicate whether this reasoning was correct or not and explain why. Codes were created for this item in order to examine more closely how students reacted when presented with the incorrect idea that random sampling leads to causation. This would potentially reveal a "pervasive fundamental misconception" (Derry et al., 2000, p. 758) between random sampling and random assignment. The code I-LAB13-RSCC was created to record when students were incorrectly agreeing that random sampling in the study design would enable causal claims. There were two ways students could correctly explain why the colleague was wrong. One way was to point out that random sampling only helps with making generalizations (C-LAB13-RSGEN). The other way was to point out that random assignment is needed for causal claims (C-LAB13-RACC).

Lab question 14 presented students with a hypothetical colleague who wanted to avoid peanuts during pregnancy based on the results of a study showing that women who eat peanuts during pregnancy are significantly more likely to have infants with peanut allergies. This study used neither random sampling nor random assignment. Students were asked to give this colleague advice on her decision based on the study design. This item was also coded separately because it was of interest to see whether students would bring up the lack of ability to make generalizations and/or the lack of ability to make causal claims based on the study design, based on prior research findings. For example, Derry et al. (2000) previously found in their interviews that students were more likely to bring up issues of sampling than of assignment to groups (partly because sampling had been learned most recently in their curriculum). Also, Groth (2006) previously found that high school

students being interviewed about how to design a study did not bring up experimental design when it was relevant.

Two codes were created for question 14 to indicate a correct critique of the study based on study design: C-LAB14-NOCC (indicating the lack of ability to make causal claims) and C-LAB14-NOGEN (indicating the lack of ability to generalize). In addition, it was anticipated that students might incorrectly use only the low *p*-value, without considering study design, to state that their colleague should avoid peanuts (code I-LAB14-PVAL). This type of response was anticipated because on previous assessment data from CAOS (delMas et al., 2007) and GOALS (Sabbag, 2013), many students incorrectly indicated that a statistically significant correlation establishes a causal relationship between variables, even in a study with no random assignment. Moreover, it was also predicted that students would make a recommendation to their colleague based on factors not related to the study design or results, such as on their own contextual knowledge of peanut allergies (code I-LAB14-NOSD). Researchers such as Wroughton et al. (2013) have hypothesized that students may have a tendency to answer statistical questions based on whether a conclusion agrees with their opinion on the context. Derry et al. (2000) also found that students often relied on non-statistical arguments to answer questions about claims of studies, and fault findings because of inconsistencies with their own prior beliefs. Therefore, it was predicted that students might give advice to the hypothetical colleague based on their own beliefs about peanut consumption and allergies, or other non-statistical arguments.

Development of codes specific to group quiz: According to observer notes during the group quiz, students often had problems judging whether a headline was making a

generalization and/or a causal claim. Therefore, when coding items on the quiz that were related to headlines based on statistical studies, the codes I-QUIZ-HGEN (difficulty recognizing a generalization from a claim) and I-QUIZ-HCC (difficulty recognizing a causal claim) were used. These codes appear in Table 3.7 below. Two of the three quiz contexts involved potential headlines making conclusions from statistical studies. The codes below were used when coding behaviors for these two contexts.

Table 3.7
Behaviors used for qualitative analysis coding, specific to group quiz

Code label	Coded behavior	Source
I-QUIZ-HGEN	Difficulty recognizing whether a headline is making a generalization	Classroom observations
I-QUIZ-HCC	Difficulty recognizing whether a headline is making a causal claim	Classroom observations

3.10 Chapter summary

A two-and-a-half-week unit about study design and conclusions was developed to help students learn the distinction between random sampling and random assignment, and conclusions that can be made from each. The unit included four activities, one group quiz, and one lab assignment. The lessons were implemented in four sections of an undergraduate introductory statistics course, and observed by the researcher and a co-observer. The Inferences from Design Assessment (IDEA) was developed as a forced choice assessment and administered as a pretest and posttest. Data from IDEA were analyzed quantitatively, and student answers from the lab assignment and group quiz were analyzed qualitatively. The results of the class observations, and the quantitative and qualitative analyses of the assessments are presented in the following chapter.

Chapter 4

Results

4.1 Introduction

In order to examine introductory statistics students' understanding of study design and conclusions, a study design unit was implemented, lessons were observed, and three assessments were administered (one forced-choice and two constructed response). This chapter first describes the results from the class observations for each activity. Then, results from the IDEA test, including changes from pretest to posttest, are presented. Finally, the chapter describes results from the qualitative analysis of the group quiz and lab assignment that were completed near the end of the unit.

4.2 Results from class observations of activities

This section provides the results from class observations of the activities in the study design unit. The three in-class sections were observed by the researcher and a co-observer for the five days of the unit. The purposes of these class observations were to (1) document how the lessons were implemented, including the extent to which the lessons were implemented as according to the lesson plan (fidelity), and to (2) explore how students reacted to the lesson, including areas of perceived understanding and areas of perceived difficulty. The researcher observed the online section by reading the activity discussion forums and group summaries. The purpose of this was to see how students discussed the activities, exploring areas in which they appeared to understand concepts and areas in which they appeared to have difficulty.

For the in-class sections, a lesson plan was shared and discussed with instructors prior to the lesson implementation. An observation form including a checklist for each

element of the lesson plan was filled out by the observers. For the online section, the lesson plans were given to the instructor so that he was aware of the main points to emphasize throughout discussion and in wrap-up videos or documents. The online instructor monitored all discussions and intervened in discussion groups for all activities, except for "Murderous Nurse" which happened during the same week as "Strength Shoe."

In this section, findings from these in-class and online observations are highlighted for each of the four activities. First, the classroom observation checklists are described. Then, the implementation of the lessons is discussed, as the first purpose of the observations was to examine fidelity of lesson implementation. For each activity, a table is provided summarizing the results of the observation checklist. Similarities and differences between sections in the implementation of the activities are discussed. Next, observation notes from each activity are summarized to provide information relevant to students' understanding of the concepts being taught, as the second purpose of the observations was to explore students' development of understanding.

4.2.1 Classroom observation checklists

The classroom observation checklists used by the observers are found in Appendix E. The observer placed a check mark by each element that (s)he observed during the lesson. For all lesson plans in this study, the elements labeled "L" represent elements of the lesson plan intended to be addressed in large-group, or whole-class, discussion. The elements labeled "S" represent potential student questions or issues that the researcher anticipated could come up during the activity time, and each potential issue had one or more suggestions for how the instructor could respond. The following subsections present tables summarizing the lesson plan elements and how many times each element was checked.

There were two observers per section; therefore, the maximum number of times each element could have been checked for each section is two. Some elements of the lesson plan were suggested ways to address potential student issues. If an issue did not arise, the potential suggestions corresponding to that issue are omitted from the table, as they were not checked.

4.2.2 *Sampling Countries* activity

The first activity, "Sampling Countries" (Appendix B1), focused on methods of taking a sample from a population and concepts of biased and unbiased sampling methods. In the activity, students first were asked to come up with a sample of countries they believed to be representative of the world. As a class, they plotted the mean life expectancies from their convenience samples. Then, they went on to take random samples and plot each of those means. The goal of this activity was for students to learn how a convenience sampling method may tend to produce biased estimates, while random sampling tends to produce unbiased estimates of the parameter.

Sampling Countries: Classroom Observation Checklist

Table 4.1 below summarizes the results of the observation form checklist for "Sampling Countries" (Appendix E1). Both the researcher and the co-observer in each section agreed that all required lesson elements had been covered, except there were some inconsistencies in section 3 between the researcher and the co-observer's checklist at the beginning. This is because the co-observer arrived late and was not there for the activity introduction. Also, only one observer checked off some of the suggested (not required) questions for large group discussion, because they were discussed in various small groups during activity time, rather than in the final large group discussion.

Table 4.1
Summary of observation checklist results for "Sampling Countries" activity.

Element from lesson plan[a]		Section 1	Section 2	Section 3
L1.	**Instructor briefly introduces activity[b]**	2	2	1
L2.	**Instructor gives 20-25 minutes for first part**	2	2	1
S1.	Students ask what is meant by "representative.	1	0	0
	S1A. Instructor asks students to come up with snapshot of countries.	0	0	0
L3.	**Instructor plots averages on *TinkerPlots***	2	2	2
L4.	**Instructor asks students to continue activity.**	2	2	2
S2.	Plot of students' samples actually centered at parameter.	0	0	0
S3.	Students ask why random samples are smaller than their convenience samples.	0	0	0
S4.	Students ask what "similar" means when comparing sample statistics.	0	0	0
L5.	Difference between sample and population?	0	2	1
L6.	Difference between statistic and parameter?	2	2	1
L7.	Center of plot of convenience sample statistics	0	0	2
L8.	**Is naming countries a biased sampling method?**	2	2	2
	L8A. Why/why not?	2	2	2
L9.	**What does it mean for sampling method to be unbiased?**	2	2	2
L10.	**Is random sampling an unbiased method?**	2	2	2
	L10A. How can you tell based on plot?	2	2	2
L11.	**Question #20 (larger biased sample vs. smaller random sample)**	2	2	2
L12.	**In real life, only have one sample.**	2	2	2
L13.	**Need to use unbiased sampling method**	2	2	2

[a] Elements in bold indicate required parts of the lesson plan. Elements not bolded indicate suggested components or suggestions for potential issues that might arise and how to address them.
[b] Elements are numbered. Those that begin with *L* indicate large group discussion components of the lesson, whereas those that begin with *S* indicate small group discussion suggestions.

Sampling Countries: Classroom instructor implementation

Based on the observer notes, although the in-class instructors included all required components of the large group discussion, each instructor introduced the unit a bit differently. The instructor of sections 1 and 3 gave a brief example of how researchers may be interested in finding out something about the population of University of Minnesota students, but cannot survey all of them. She asked the students what they would do in this

case, and led a very brief discussion on ways to sample students. The instructor of section 2 placed three questions about study design on the board:

(1) How were the participants/subjects selected to be in the study?
(2) Once selected (if comparing groups), how were the subjects "assigned" to their groups?
(3) Is a larger sample always better?

Then, she indicated that the class would be focusing on the first and third questions this time and the second question later on.

The instructors also addressed all of the required wrap-up questions, although they did not have sufficient time to address most of the suggested wrap-up questions. All in-class instructors handled the wrap-up similarly by leading a large group discussion, but had different styles of calling on students to answer. The instructor of section 2 tended to call on volunteers who raised their hands, and a variety of students participated. The instructor of sections 1 and 3 tended to call on specific students or tables of students to answer.

As noted in the lesson plan, instructors emphasized that in real studies, only one sample is taken. However, the instructors emphasized this idea in slightly different ways. The instructor of sections 1 and 3 projected the plot of the population and showed that there are some countries with very low life expectancies. She pointed out that just by chance, it is possible to get an unusually low sample mean if many of these countries happen to be chosen in the random sample. Then, she showed the plot of the 200 sample means from the random sampling and showed that unusual values are rare. The instructor of section 2 drew a picture of the many randomly sampled means on the board, and then drew another plot with just one sample mean, displaying how in real life only one of these dots is visible. She

talked about having "faith" that random sampling is an unbiased method, and that we know this method will not tend to over- or under-estimate the parameter.

The "Sampling Countries" activity was designed to address the potential misconception that larger samples are always better, regardless of sampling method. As requested in the lesson plan, all instructors emphasized in their wrap-ups that smaller, random samples are better than large, convenience samples (even though the ideal situation would be to have a large, random sample). The instructor of section 2 spent the most time on this topic, giving students an example with a different context. She presented students with the scenario of estimating the average income of University of Minnesota graduates, with two possible samples: a larger sample of students from an alumni event, and a smaller sample of students randomly chosen from the registrar's records. After allowing students to discuss, she asked for a show of hands and found that almost all students correctly preferred the smaller random sample from the registrar. The instructor of sections 1 and 3 did not provide a different context example as the instructor of section 2 did. However, when she asked students about this concept in the wrap-up, students correctly indicated that a smaller, random sample would be preferable to the larger, biased one. Table 4.2 summarizes some of the differences in methods used by the in-class instructors during large-group discussion of "Sampling Countries."

Table 4.2
Summary of methods used during large-group discussion of "Sampling Countries" activity

Method	Section 1	Section 2	Section 3
Introduced unit by leading discussion on how to take samples	X		X
Introduced unit by writing questions about study design on the board (and revisited questions throughout unit).		X	
Called on specific students (or groups) to answer questions	X		X
Asked for student volunteers to answer questions		X	
Used *TinkerPlots* to randomly select students to respond to questions			X
Pointed out "outliers" in population that could lead to unusual sample means	X		X
Drew picture of one sample mean on the board showing that in real studies, only one sample is taken		X	
Provided example, using different context, of comparing a larger, biased sample with a smaller, random sample		X	

Note. An "X" in a cell indicates the instructor of that section used the corresponding method.

Sampling Countries: Online instructor implementation

Due to the asynchronous nature of the online class, a large group wrap-up discussion was not possible in this section. Instead, the online instructor addressed the required wrap-up questions in a four-and-a-half minute video. In the video, he first went through the *TinkerPlots*^{TM} logistics on sampling randomly and showing how the plots of convenience sample means and of random sample means differed. He showed that the process of random sampling will, on average, provide the correct estimate, even if the sample statistic is not exactly the same as the parameter each time. He added that the variability in sampling is taken into account when *p*-values and confidence intervals are computed. In the video, the instructor also mentioned that in real life, only one sample is taken, but random sampling helps the sample to be representative of the population, allowing for unbiased estimates and generalizations being made to the population.

Sampling Countries: In-class addition of discussion on making generalizations

Although the online instructor addressed generalization in his wrap-up, the in-class instructors and researcher discussed after the "Sampling Countries" activity that they had focused mostly on biased vs. unbiased sampling methods, and had not addressed what it meant to make generalizations. Therefore, it was decided that they take some time at the beginning of the following class to address why unbiased sampling methods can lead to generalizations to a population. The instructor of sections 1 and 3 spent about 15 minutes of the following class period on this topic. She projected a plot of the population and the sample (see Figure 4.1), explaining that researchers typically take a sample because they do not see the population. She led a brief large group discussion about how, even though in reality only the sample mean is seen, an unbiased sampling method allows one to conclude that the population mean is somewhat similar.

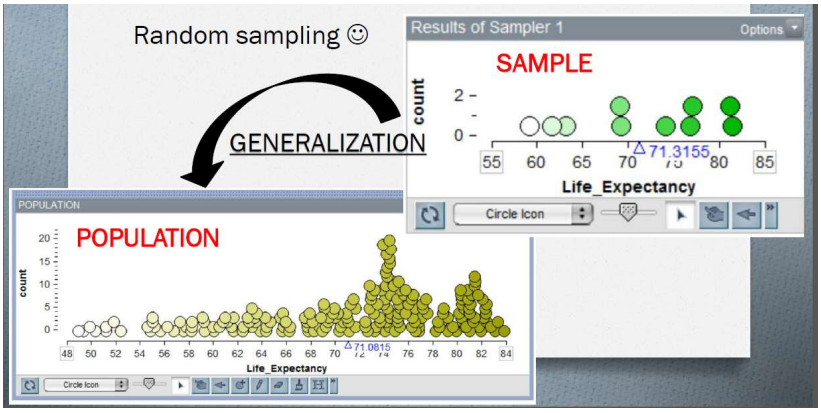


Figure 4.1. Slide shown by one instructor about using a sample to generalize to a population.

The instructor of section 2 spent about 25 minutes of the following class period addressing generalization. She asked students to read the last paragraph of the activity which talked about generalization, and asked them to reflect on how they selected their 20 countries. She led a large group discussion about how non-random samples make the population more difficult to define – for example, the convenience samples might allow students to generalize to "countries that come to the minds of EPSY 3264 students." Although the instructor of sections 1 and 3 did not lead a discussion about generalizing to limited populations from convenience samples, she did mention that researchers can still use their data and results while being careful about to whom they generalize.

Sampling Countries: Observations of activity discussions

One purpose of the observations was to examine how students went through the activities, including potential obstacles or questions. As shown in Table 4.1, the issues and questions that the researcher anticipated might arise during small group activity time were not observed often. However, instructors did bring up some of those issues as a part of large group discussions. For example, it was anticipated that students might ask the instructor how to come up with a "representative" sample of 20 countries. Students were not observed asking this, but during the wrap-up, the instructor of section 2 asked students to discuss how they had thought of their "representative" sample. Also, it was anticipated that students might ask the instructor why their convenience sample had been size 20 and the random samples had been of size 10. Instead, the instructor addressed this question in the wrap-up when talking about whether it is better to have a larger convenience sample or a smaller random sample.

Instead of the anticipated issues for the activity, other issues arose. The first was confusion about the terminology of the words "parameter" and "statistic" when students were asked whether their sample mean life expectancy of 20 countries was a parameter or a statistic. Although the terms had been briefly defined in a short reading in the activity prior to those questions, the observers overheard many students asking what the terms meant. In all three sections, the in-class instructors tried to scaffold students with questions, such as asking them the difference between sample and population, and asking them what they were trying to estimate. After instructors' scaffolding, most of the in-class students observed successfully identified the parameter as the average life expectancy of the population and the statistic as the average life expectancy of the sample.

Online, confusion about "parameter" and "statistic" was not observed in the discussion boards, but the online students' group discussion questions did not ask them to identify whether their sample means were parameters or statistics. Still, the online instructor explained the difference between "parameter" and "statistic" in the wrap-up video he made. It is unclear whether the online students understood this distinction after the activity, as the instructor scaffolding happened as part of the activity wrap-up video and students did not discuss the activity after this.

Observations revealed that students also struggled with some of the other vocabulary terms, using colloquial meanings instead of statistical meanings. For instance, students sometimes used the word "random" to mean haphazard. When one instructor asked students to think about how they selected their 20 countries at the beginning of the activity when they were asked to think of countries that were representative, one student replied "randomly." Several students in the online section also reported that they chose

their countries "randomly" even though it was clear that they had purposefully chosen them for different reasons. Also, the colloquial meaning of the word "bias" was used by a student attempting to answer a question posed by the instructor about how we can know that a sampling method is unbiased. The student brought up that to avoid bias, it was important to evaluate who is collecting the data (e.g., researcher bias). Instructor responses to these vocabulary issues were to ask students more questions, such as asking whether the 20 countries really were taken at random, and asking what bias meant in a statistical sense.

Another problem that some students faced during the activity involved reasoning about repeated sampling. Although students had already had experience with repeated trials in randomization tests, some of them were observed having problems with predicting what a plot of 200 sample statistics would look like using an unbiased sampling method. Some in-class students predicted that their plot of sample statistics would be skewed left, because they thought most of the countries had higher life expectancies. This shows potential confusion between sample means and individual data points. Other students tended to focus on predicting the shape and variability of the plot, but omitted predictions of the center, which is more pertinent to the concept of bias. Although online students were not asked to post their predictions of what a plot of repeated sample statistics would look like, some of their answers revealed possible misunderstandings about the notion of repeated sampling. When students were asked to reason about whether random sampling tended to produce unbiased estimates, many students wrote on the discussion boards comments such as "more samples are needed" or "more trials are needed." Most of the intervention by the instructor in the online discussion boards involved explaining that in reality, a single study rarely gets more than one trial. In the wrap-up video, the instructor addressed the idea that even though

only one sample is taken, it should be taken with a sampling method that tends to provide unbiased estimates.

In summary, the following issues arose, none of which were specifically addressed in the lesson plan:

- Difficulty defining the terms "parameter" and "statistic"
- Use of colloquial meanings of the terms "random" and "bias
- Failing to distinguish between distributions of individual case values and distributions of sample statistics
- Problems understanding the notion of repeated sampling

Despite these observed issues with students' reasoning, the observations revealed overall that there were many instances of students reasoning correctly about sampling methods and bias, both in class and online. However, the instructors and observers shared in post-activity feedback that there were two concepts that needed to be emphasized more: (1) why random sampling is an unbiased method (rather than just trusting that random sampling would be unbiased), and (2) what it means to generalize to a population.

4.2.3 *Strength Shoe* activity

The second activity, "Strength Shoe" (Appendix B3), allowed students to explore how random assignment tends to balance out confounding variables in the long run, and does this better than purposefully assigning participants to groups. Prior to this activity, students were required to complete the reading "Establishing Causation," which introduces ideas of explanatory and response variables, confounding, and random assignment. Some

extra questions for discussion were suggested in the lesson plan in case there was additional time, but there was no extra time in any of the in-class sections.

Strength Shoe: Classroom Observation Checklist

Table 4.3 shows the observation checklist indicating the number of times each lesson plan element was checked for each section. In the lesson plan, there were some extra suggested questions to address in large-group discussion if time allowed. These were not addressed due to lack of time, so they are omitted from the table. The full lesson plan checklist for this activity is found in Appendix E2. For the most part, there was agreement between the two observers that almost all required parts of the lesson were addressed. For section 1, the co-observer did not check off some of the large-group discussion questions. However, according to the observation notes, the wrap-up questions that were checked on the checklist by the researcher were displayed on PowerPoint slides for students to discuss, and the instructor's large group discussion addressed the ideas in these.

Table 4.3
Summary of observation checklist results for "Strength Shoe" activity.

	Number of times element was checked (2 observers per section)		
Element from lesson plan[a]	Section 1	Section 2	Section 3
L1. Instructor briefly introduces activity[b]	2	2	2
L1A. Instructor asks students if they have heard of StrengthShoes	0	2	1
L2A. Instructor asks students about anecdotal evidence in this context.	0	2	0
L3. Instructor asks students to work on activity in groups.	2	2	2
S1. Students struggle with question about why random sampling is preferable.	0	2	0
S1A. Instructor asks what student(s) learned about random sampling in last activity	0	0	0

	Number of times element was checked (2 observers per section)		
Element from lesson plan[a]	Section 1	Section 2	Section 3
S2A. Instructor asks what kinds of conclusions can be made when random sampling is used	0	1	0
S2. Students think "balanced" means 50% in each group	0	1	0
S3. Students have trouble judging whether groups are "roughly equivalent"	2	2	0
S3A. Instructor asks if groups are more or less equal, or very different	1	1	0
S4. Students struggle to predict what plot of many differences will look like with random assignment	1	0	2
S4A. Instructor asks students to run sampler more	0	0	0
S4B. Instructor asks students to predict what happens if sampler is run 100 more times	0	0	0
S5. Students struggle to reason about plot center at 0	1	0	1
S5A. Instructor asks what each dot represents	0	0	0
S5B. Instructor asks what a dot of 0 means	0	0	0
S5C. Instructor asks why it makes sense that the distribution is centered at 0	1	0	0
S6. Students struggle to answer whether random assignment allows for a cause-and-effect conclusion	0	0	2
S7. Students skeptical about random assignment balancing out confounding variables	0	0	2
S7A. Instructor asks if perfect balance is possible in single trial	0	0	0
S7B. Instructor asks about balance in the long run	0	0	1
S7C. Instructor asks: If groups are balanced, is it likely that confounding variables are responsible?	0	0	0
S8. Students struggle with whether or not they can generalize	0	0	1
S9. Students confuse "generalization" with "causation"	0	0	1
S10. Students think only small sample size inhibits generalization	0	0	0
L4. What is the treatment variable?	0	0	0
L5. What is the response variable?	0	0	0
L6. What does it mean to make a causal claim?	1	2	2
L7. What is a confounding variable?	0	2	2
L8. How can confounding variables limit ability to make causal claims?	1	2	2
L9. Is purposeful assignment to groups a good idea?	1	2	2
L9A. Why/why not?	1	2	2
L10. Is it possible to get perfect balance in one single random assignment?	0	2	0

	Number of times element was checked (2 observers per section)		
Element from lesson plan[a]	Section 1	Section 2	Section 3
L11. Why were plots of differences centered around 0?	1	2	2
L12. Does random assignment tend to balance out confounding variables?	2	2	2
L13. Why can we make cause-and-effect conclusions with random assignment?	1	2	2
L14. In real life, we do not perform repeated random assignments	0	2	0
L15. In reality, there is only a single random assignment	0	1	0
L16. The method of random assignment needs to be one that tends to balance out confounding variables	0	2	0
L17. What is the difference between random assignment and random sampling?	2	0	2
L18. Did this study use random sampling? How would this affect our potential conclusions?	0	0	0

[a] Elements in bold indicate required parts of the lesson plan. Elements not bolded indicate suggested components or suggestions for potential issues that might arise and how to address them.
[b] Elements are numbered. Those that begin with *L* indicate large group discussion components of the lesson, where those that begin with *S* indicate small group discussion suggestions.

Strength Shoe: Classroom instructor implementation

Based on the notes taken by the observers, the in-class instructors each introduced the activity in all three sections a bit differently. The instructor of section 2, having had personal experience with similar shoes called "Jump Soles," showed a YouTube video about them and told a personal story about wearing them. She then asked students: "If someone were to wear these shoes and their jumps got better, would you conclude that the shoes work?" After students had time to discuss, a few of them reasoned correctly in large group discussion that other factors such as genetics and athletic ability may affect jumping, and one person's results are not enough. The instructor of sections 1 and 3 spent less time discussing the specific Strength Shoe context in her introduction, and more time on the idea of random assignment. Sections 1 and 3 began with a 3-minute pop quiz on the

"Establishing Causation" reading at the beginning of class, and then a brief discussion on what it meant to randomly assign subjects to groups. One student responded that random assignment involved taking a "random sample" of subjects and assigning them at random into groups, and the instructor clarified that random assignment can be done even if the sample itself is not random.

According to the checklists, the instructors addressed all required elements of the lesson plan, except for some final take-away points that instructors were supposed to make at the end. These take-away points involved the idea that in real studies, only one sample is taken, and it is necessary to use a method that tends to balance out confounding variables. The instructor of section 2 addressed these points, but the instructor of sections 1 and 3 instead chose to focus on discussing the difference between random sampling and random assignment.

All three in-class sections spent part of the beginning of this class period (Day 2 of the unit) with some additional wrap-up of the "Sampling Countries" activity and what it meant to generalize. After giving students time to work on the "Strength Shoe" activity, the in-class instructors all postponed their main wrap-up until the following class period (Day 3). The instructor of section 2 announced that she would give students 10 minutes to finish at the beginning of the following class period. The instructor of sections 1 and 3 asked students to finish the activity at home if necessary, and led a very brief preliminary wrap-up lasting less than five minutes. This brief wrap-up involved a large group discussion about the purpose of random assignment, emphasizing that random assignment balances out confounding variables so that the only variable that is different between the groups is the shoe. She also briefly clarified the distinction between random sampling and

random assignment, mentioning that random sampling is how to select the subjects in the first place, and random assignment to groups happens after subjects are already selected.

In all three sections, the wrap-up of the "Strength Shoe" activity occurred on Day 3 of the unit, before beginning the "Murderous Nurse" activity. This wrap-up time in each in-class section involved going back and forth between small-group and large-group discussion of the main wrap-up questions. The instructor of sections 1 and 3 placed sets of questions on slides and asked students to discuss each set of questions, while the instructor of section 2 went back and forth more frequently between small-group and large-group discussion, asking students to discuss one question at a time. After the main "Strength Shoe" activity wrap-up questions, the instructors went on to discuss the differences between random sampling and random assignment. Table 4.4 summarizes some of the differences in methods used by the in-class instructors during large-group discussion of "Strength Shoe."

Table 4.4
Summary of methods used during large-group discussion of "Strength Shoe" activity

Method	Section 1	Section 2	Section 3
Pop quiz on the "Establishing Causation" reading	X		X
Introduction of activity context: What are Strength Shoes?		X	
Introductory discussion about anecdotal evidence on the effectiveness of Strength Shoes		X	
Introductory discussion focused on what it means to randomly assign	X		X
Wrap-up split into large-group and small-group discussion time of key questions	X	X	X
During wrap-up, asked students to discuss a set of projected questions at a time	X		X
During wrap-up, asked students to discuss one question at a time		X	
Wrap-up discussion occurred during Day 3 of unit	X	X	X
Gave students 10 minutes to finish activity at the beginning of Day 3 of unit		X	
Asked students to finish activity at home at the end of Day 2	X		X
Discussed distinction between random sampling and random assignment	X	X	X
Confirmed students' correct answers during large group discussion	X		X
Sought student consensus during large group discussion, without confirming correct answers.		X	

Note. An "X" in a cell indicates the instructor of that section used the corresponding method

Strength Shoe: Online instructor implementation

In the online class, the instructor addressed the wrap-up questions in a few paragraphs of about 900 words total. He emphasized that although perfect balance isn't possible in one trial, random assignment tends to balance out "lurking variables," on average. Because of this, if a statistically significant result is found, we can attribute it to the explanatory variable that was randomly assigned. In the wrap-up, the online instructor called the ability to make causal claims "internal validity." The in-class instructors did not use the term "internal validity" and they also used the term "confounding variables" rather than "lurking variables."

The online instructor focused mostly on ideas of random assignment, confounding, and causation in his written wrap-up, and spent less of his wrap-up on the distinctions between random sampling and random assignment than the in-class instructors. However, he did address random sampling, saying that even if the subjects were randomly assigned to wear Strength Shoes or ordinary shoes, random sampling would be necessary for having a strong statistical argument that Strength Shoes help people jump farther. He also mentioned that in reality, it is difficult to have both random sampling and random assignment, but when it is crucial to attempt to make a causal claim in the study, having random assignment is more important. The online instructor gave an example regarding medical trials, which often emphasize random assignment, because the main question is to conclude whether drugs or treatments will cause improvement.

Strength Shoe: Observations of activity discussions

Unlike with the "Sampling Countries" activity, most of the issues and questions that the researcher anticipated might arise during the small-group activity arose in at least one section of the class. In the activity, students first examined a purposeful assignment of groups that balanced out subjects with respect to sex and height, with each group being composed of 4 males and 2 females, and the average heights being approximately equivalent. One anticipated issue was that students would claim that the groups were not balanced because they were not 50% male and 50% female. Only one in-class student group, in section 2, was observed claiming that the groups were not balanced with respect to sex, because the subjects in each group were not half male and half female. Online, this issue was prevalent, with many students claiming on the discussion boards that the groups were not balanced with respect to sex because males were overrepresented. Another

anticipated issue that arose occasionally in sections 1 and 2 was difficulty judging whether groups were "roughly equivalent" with respect to certain variables. As recommended in the lesson plan, both instructors asked groups to think about whether groups were more or less equal, or very different.

As previously anticipated, some students had difficulties predicting what a plot of differences would look like for many random assignments, especially in the online class. In the online discussion boards, instead of predicting the plots would be centered at 0, many students instead gave a range of values, such as "between -1 and 1" or "between -2 and 2." In section 3, one group of students had successfully predicted and observed that the differences in mean heights would be centered at 0 for many random assignments, but had problems making a prediction for the center of the plot of differences in percent of subjects with the X-factor. The instructor observed this difficulty and encouraged the students to apply the same reasoning they had just used for examining the height variable.

While many students were observed correctly reasoning that random assignment tends to balance out confounding variables when looking at their plots of differences, some students still questioned, at the end of the activity, whether random assignment (and a significant difference) would allow researchers to conclude that the type of shoe caused the difference in jumping ability. This issue had been anticipated because it has been documented in the literature that students tend to be skeptical about the effectiveness of random assignment to balance out groups (e.g., Sawilowsky, 2004). For example, one group of in-class students in section 3 claimed that the "X-factor" might still be a confounding variable after the random assignment. When the instructor observed this, she pointed the students to the plot they had made of differences in proportions of subjects with

the "X-factor" and asked them whether random assignment tended to balance out groups with respect to this variable. Online, many students were also skeptical about the effectiveness of random assignment, but mainly due to the small sample size. The instructor addressed this concern by discussing that statistical methods account for sample size, and when the sample size is small it is harder to get a small *p*-value. A few students were critical of the design of the randomized comparative experiment itself, saying that it would instead be better to have subjects jump once with each type of shoe (thus suggesting a matched pairs design).

Near the end of the activity, students were asked to conclude whether one could generalize results of the study to all athletes. In class, some students asked the instructors about this question because the activity handout first described a previous study done with 12 intercollegiate track athletes, and then went on to have students consider a hypothetical study recruiting 12 of their friends to participate. The answer to the generalization question was the same regardless of which sampling option they considered, but students were confused about which sampling option was being referred to in the question. Some students cited the small sample size as the reason why the results were not generalizable, especially in the online class.

Also, in the online class, many discussion posts contained incorrect answers to the generalization question because students were not interpreting the question correctly. Some students were equating the word "generalize" with the phrase "make causal claims," while others interpreted the question as asking whether one could make causal claims from the study with purposeful assignment. Various in-class students were observed correctly discussing the lack of random sampling, and referring to the previous discussion they had

just had about this topic in class. The online students who did interpret the question correctly noted that the sample overrepresented males, and that the subjects were not randomly sampled.

In summary, the following issues arose in this activity, most of which were addressed in the lesson plan:

- Believing that "balanced" groups means 50/50 balance, rather than approximately the same distribution of outcomes in both groups
- Difficulty predicting what a plot of differences will look like after many random assignments
- Stating that confounding variables should be a major concern, even after random assignment (i.e. not believing in the effectiveness of random assignment)
- Emphasizing small sample size over study design method
- Confusing the terms "generalization" and "make causal claims"

Despite these obstacles, many student groups were observed correctly using their plots of differences to conclude that random assignment helps to balance out groups with respect to confounding variables. They appeared to have more difficulty, however, applying this reasoning to discuss whether one could make causal claims using data from a study which uses random assignment.

Strength Shoe: In-Class addition of discussion on distinguishing between random sampling and random assignment

The instructors in all three sections discussed the differences between random sampling and random assignment after the main "Strength Shoe" activity wrap-up. The

instructor of sections 1 and 3 began her wrap-up by asking her students the difference between random sampling and random assignment. When student volunteers answered, she confirmed when they answered correctly. She emphasized that random sampling and random assignment happen as part of the study design, *before* data are collected. The instructor of sections 1 and 3 also emphasized the importance of being able to explain *why* random sampling allows for generalizations (the importance of representative samples), and *why* random assignment allows for causal claims (the balancing out of confounding variables).

In contrast, the instructor of section 2 led a much longer discussion about the distinction between random sampling and random assignment after addressing the main wrap-up questions in the "Strength Shoe" activity. She asked students some "yes or no" questions and asked them to give a thumbs-up for "yes" and a thumbs-down for "no." When she asked students if random assignment allows them to generalize, many students gave a thumbs-up. Immediately thereafter, she asked whether random assignment allows for causal claims, and fewer students gave a thumbs-up. When the instructor asked if random assignment allowed for both generalizations and causal claims, one student referred to the "Establishing Causation" reading and asked whether random assignment is the correct design for causal claims. Rather than answering, the instructor turned student questions back to the whole class and asked what others thought. After 30 minutes of wrap-up discussion, students could not come to a consensus about whether random assignment allowed for generalization, causation, or both. The instructor then had to stop the discussion and explain random assignment and random sampling in a mini-lecture more fully, in order to allow enough time for the next activity, "Murderous Nurse."

4.2.4 *Murderous Nurse* activity

The third activity, "Murderous Nurse," had students carry out a randomization test for a difference in proportions, which they had done various times previously in the course. This time, however, they were asked to consider the scope of inferences that could be made based on the study design. The proportion of shifts in which a death occurred when the nurse Kristen Gilbert was working was compared to the proportion of shifts in which a death occurred when she was not working. In the study referenced by this activity, neither random sampling of shifts nor random assignment of shifts occurred. Prior to this activity, students were assigned the "Scope of Inferences" reading, which distinguished between random sampling and random assignment, discussing the types of conclusions that could be made from each.

Table 4.5 below shows an abbreviated observation checklist with the number of times each lesson plan element was checked for each section. The full lesson plan checklist for this activity is found in Appendix E3. The instructors covered nearly all of the required main points, although due to the longer "Strength Shoe" wrap up in section 2, the instructor of that section did not spend as much time as the instructor of the other sections on the "Murderous Nurse" activity or wrap-up.

Table 4.5
Summary of observation checklist results for "Murderous Nurse" activity.

Element from lesson plan[a]		Section 1	Section 2	Section 3
L1.	Instructor mentions return to randomization tests[b]	2	2	2
L2.	Instructor mentions study design will now be considered	2	2	2
L3.	Instructor asks students to work through activity in groups	1	0	2
L4.	Instructor asks students to check #1-6 with others	1	2	1

Element from lesson plan[a]	Section 1	Section 2	Section 3
L5. Instructor mentions if students done early, can search for information on Kristen Gilbert online	2	2	0
S1. Students unsure on explanatory/response variables	2	2	2
S1A. What variable do we want to predict here?	1	2	0
S1B. Which variable can help us predict it?	1	2	0
S2. Students struggle to answer what dots in plot represent	0	0	2
S2A. Instructor asks what the null model is	0	0	1
S3. Students confuse random assignment in randomization test with random assignment in data collection	1	0	1
S3A. What are you modeling in this simulation?	0	0	0
S3B. How were shifts in original data divided into groups?	0	0	0
S3C. Instructor points out difference between null model and data collection	1	0	0
S4. Students struggle to find *p*-value	1	0	0
S4A. Where is the observed result on the plot?	1	0	0
S4B. How many trials are beyond the result?	0	0	0
S5. Students struggle with study design questions	1	1	1
S5A. How were shifts sampled?	1	0	0
S5B. How were shifts assigned?	1	0	0
L6. What statistic did you collect?	0	0	0
L7. What does plot of 500 trials represent?	0	0	0
L8. Is observed difference statistically significant?	0	0	0
L9. What does it mean to be statistically significant?	2	0	0
L10. How did you answer the research question?	2	0	2
L11. How were shifts sampled?	2	0	2
L11A. What does this imply about conclusions?	2	0	2
L11B. What does it mean to generalize?	1	1	2
L12. How were shifts assigned?	2	2	2
L12A. What does this imply about conclusions?	1	2	2
L12B. What does it mean to make causal claims?	1	2	2
L12C. Alternative explanations for difference?	0	2	2
L13. What can be concluded, then?	2	2	2
L14. Could study be valuable in court?	0	2	0
L15. Is follow-up study with random assignment advisable?	2	2	2
L16. Observational studies can still be useful without random sampling or random assignment	2	0	2
L17. We *can* say observed difference unlikely to happen by chance	2	0	1
L18. Experiments ideal, but not always ethical	2	0	2

[a] Elements in bold indicate required parts of the lesson plan. Elements not bolded indicate suggested components or suggestions for potential issues that might arise and how to address them.
[b] Elements are numbered. Those that begin with *L* indicate large group discussion components of the lesson, where those that begin with *S* indicate small group discussion suggestions.

Murderous Nurse: Classroom instructor implementation

The in-class instructors began the activity on Day 3 of the unit, after concluding the wrap-up discussion of the "Strength Shoe" activity and after leading a large-group discussion about the differences between random sampling and random assignment. This discussion had taken different amounts of time in each section. Most notably, in sections 1 and 3, the "Murderous Nurse" activity began about 20 minutes after the start of class, and in section 2, the activity began about 30 minutes after the class period started. Since the instructor of section 2 had given students time to finish "Strength Shoe" and led a longer discussion about random sampling and random assignment, there was less time than in the other sections to complete the "Murderous Nurse" activity.

In the activity, students were not given explicit instruction on how to set up the model in *TinkerPlots*™. Since students already had experience with tests for differences in proportions, the instructors recommended letting the students set up the model themselves. In sections 1 and 3, the instructor briefly went over how to set up the model in the wrap-up discussion, but did not spend much time on this because most of her students in both sections had figured this out. In sections 1 and 3, the instructor had more large-group discussion time to dedicate to the questions about generalization and causation, and focused less on how to set up the model. In section 2, some students had difficulty setting up the sampler in *TinkerPlots*™, so the instructor decided to interrupt the class in the middle of the activity and lead a brief large group discussion about how to set up the model and find the *p*-value. The wrap-up questions about scope of inferences were planned for discussion after students had finished the activity, but a student brought up random assignment during this interruption period while the class was talking about the model. This prompted some

large-group discussion about generalization and causation before most students got to the scope of inferences questions on the activity. When students finally went back to work on the activity, there were 8 minutes of class left. During the last minute of class, the instructor briefly asked the class what was needed for generalization and what was needed for causation. Some students were overheard saying "randomization," but there was not enough time to finish this discussion. The instructor of section 2 then decided that she would lead a brief discussion on scope of inferences before the group quiz during the following period.

In all three sections, the in-class instructors addressed almost all of the required wrap-up questions, although they did not always ask these questions verbatim. For example, in section 3, rather than asking how the shifts were sampled and what this implied about conclusions, the instructor asked the class whether this study could be used to make generalizations to all shifts, and why or why not. In section 2, the question about whether or not a follow-up study could be conducted using random assignment was brought up by a student, rather than the instructor, when the student suggested that random assignment could enable causal claims, but would be "morbid." The instructor of sections 1 and 3 led a discussion about why the lack of random sampling limited generalizations, and why the lack of random assignment limited causal claims. She also asked questions to attempt to reveal possible misunderstandings about the distinction between random sampling and random assignment, such as whether random sampling would enable them to conclude that Gilbert caused the deaths.

At the end of the lesson plan, there were some suggested discussion points about how observational studies without random sampling can still be useful, and although

experiments are ideal for making causal claims, they are not always ethical. The instructor of sections 1 and 3 addressed these questions at the end of her wrap-up, but the instructor of section 2 ran out of time during this class period.

Table 4.6 summarizes some of the differences in methods used by the in-class instructors during large-group discussion of "Murderous Nurse."

Table 4.6
Summary of methods used during large-group discussion of "Murderous Nurse" activity

Method	Section 1	Section 2	Section 3
Began activity about 20 minutes into class period	X		X
Began activity 30 minutes into class period		X	
Class interrupted during activity to go over model setup in a large group		X	
Allowed students to finish activity before beginning wrap-up discussion	X		X
Walked around during activity and asked questions to groups who were finished	X		X
Discussion on how to set up the model was only brief, during wrap-up	X		X
Addressed generalization and causation in wrap-up, but did not ask most wrap-up questions exactly as stated in the lesson plan	X	X	X
Addressed wrap-up points about feasibility of study designs (e.g., experiment ethics) and how observational studies can still be useful	X		X

Note. An "X" in a cell indicates the instructor of that section used the corresponding method

Murderous Nurse: Online instructor implementation and discussion

In the online class, students participated in discussions about the "Murderous Nurse" activity, but unlike the previous two activities, the discussion was not monitored by the instructor. Various issues arose in the online class discussion that did not arise much, if at all, in the in-class discussion, and those issues will be described in this subsection.

While the in-class students were not observed having many difficulties with computation of the sample statistic, many online students gave incorrect calculations. Students were supposed to calculate the difference between the percentage of shifts when Gilbert was working in which a death occurred, and the percentage of shifts when Gilbert was not working in which a death occurred (100(40/257-34/1384) = 13.1 percentage points; see table in "Murderous Nurse" activity in Appendix B5). Most students computed this correctly, but some of them flipped the conditional probabilities, instead calculating the difference in the percentage of shifts when a death occurred in which Gilbert was working and the percentage of shifts when a death occurred in which Gilbert was not working (100(40/74-34/74) = 8.1 percentage points). Even students who had correctly defined explanatory and response variables calculated the wrong conditional probability. While the in-class students had the opportunity to check their answers with other students and correct their answers, the online students posted their individual answers first before checking with other students.

Many of the online students set up the model incorrectly, resulting in plots that were centered at numbers far away from 0. Thus, many were also unable to find the *p*-value or found incorrect *p*-values. This affected their ability to answer subsequent group discussion questions about making a conclusion. For example, one mistake was to conclude that there was no difference between the percent of shifts that had a death occur when Gilbert was working and when she was not working, because the plot of randomization differences was centered at 0.

Although the majority of online students correctly concluded that no causal claims or generalizations could be made, various incorrect ideas about scope of inferences

appeared in the answers. For example, some students said that because of the low *p*-value, one could conclude that Gilbert caused the additional deaths. Some students said that both random sampling and random assignment were needed to make causal claims, and others assumed that random sampling had been conducted even though the activity made no mention of this.

In one of their group discussion questions, online students discussed the usefulness of this study and whether a follow-up study using random assignment would be advisable. Most students agreed that this study could still be useful, but they varied greatly in their recommendations regarding follow-up studies. Some students suggested gathering more evidence, though they did not refer to statistical evidence. Instead, they suggested observing Gilbert to see if she was tampering with the medicine. Others suggested comparing Gilbert's data to data of other individual nurses. Some students recommended further studies using random assignment, not realizing the ethical implications of that, while others did realize the ethical implications and advised against doing follow-up studies.

In general, there were many incorrect ideas in the online forum for this activity, such as inability to set up the model and find the *p*-value, inability to reason correctly about scope of inferences, and failure to recognize ethical concerns about a follow-up experiment. Since this activity happened during the same week as "Strength Shoe," there was less time for instructor monitoring of discussions. Also, since no summary was required, students did not receive any feedback on their answers.

However, the instructor posted a general wrap-up after the activity, addressing the main wrap-up questions and points in the lesson plan in about 400 words. The wrap-up

addressed the small *p*-value and the low likelihood of observing this difference in percentages just by chance. Also, the wrap-up addressed the ethical concerns about conducting an experiment. The instructor also emphasized that although this observational data could be a useful starting point for an observation, it would not be appropriate to convict Gilbert without more causal evidence. As recommended in the lesson plan, the online wrap-up emphasized issues of data collection and scope of inferences, but did not address how to obtain the statistical model and *p*-value, despite the fact that various students had problems with this part of the activity. However, the online wrap-up cited the original study and gave more information than the in-class instructors gave about the context and how the data were used in court.

Murderous Nurse: Observations of activity discussions

Most of the issues that the researcher anticipated might come up for the "Murderous Nurse" activity did arise in a few student groups. In all in-class sections, as anticipated, students were observed asking about how to define the explanatory and response variables. They had seen definitions of explanatory and response variables in readings, and had talked about "treatment" variables prior to this unit in class, but many students needed instructor intervention to identify the variables correctly. In contrast, almost all online students defined the explanatory and response variables with no problem. This might be because the "Scope of Inferences" reading was embedded within the online activity, meaning that online students were more likely to have read it more recently than in-class students who would have completed the reading prior to the class period.

One anticipated issue was that students would have problems finding the *p*-value because the observed difference was off the plot, but students had seen very low *p*-values before. Instead, problems finding the *p*-value involved having trouble setting up the model in the first place. In sections 1 and 3, the instructor walked around and helped students when necessary, but the instructor of section 2 was running short on time and had to stop the class to help them with the model. Also, in section 2, not many of the anticipated issues on the lesson plan arose, because students did not get to many parts of the activity.

In all sections, just after computing the sample statistic, students were observed correctly reasoning that even though deaths were more likely to occur during Gilbert's shifts than during other shifts, this still did not mean that she was killing patients. Students pointed out various possible confounding variables, such as the time and length of the shifts, severity of patients seen, and number of other nurses on staff. However, after running the randomization test, some incorrect ideas arose. For example, the instructor of section 1 asked in the wrap-up: "If shifts were randomly sampled, what could we say?" and a student responded that this meant we could argue that deaths were caused by Gilbert. Some students, especially online, used the low *p*-value as evidence that Gilbert caused the deaths, without considering study design, even though earlier on in the activity they had discussed the potential for confounding variables.

One of the misconceptions anticipated by the researcher was confusing the random allocation of shifts in the randomization test with random assignment in the original study. This did in fact happen in sections 1 and 3, with some groups saying that the shifts were randomly assigned because they had randomly assigned them in *TinkerPlots*^{TM}. The instructor decided to address this misconception in her wrap-up, emphasizing that the scope

of inferences depended on how the data were collected *before* the analysis was done. In section 2, this confusion between randomization in the original study and randomization in the simulation did not arise, but this may be because students were running shorter on time and getting stuck on creating their model.

In summary, students struggled somewhat with the following issues. The first two issues were addressed in the lesson plan, and the others were not.

- Difficulty defining explanatory and response variable
- Confusing the random assignment in the original data collection with the random assignment that is done in the randomization test simulation
- Not recognizing whether random sampling was done in the original study
- Mistakes in calculating the sample difference in proportions
- Difficulty setting up a model to test for a difference in proportions to find the *p*-value
- Using the low *p*-value to justify causal claims

Despite these obstacles, many student groups in sections 1, 3, and 4, were observed correctly reasoning about the inability to make generalizations and the inability to make causal claims based on the study design. Although section 2 ran out of time, a follow-up discussion of generalization and causation occurred during the following class period.

Section 2: Pre-quiz discussion

Because section 2 did not have enough time to adequately wrap up the "Murderous Nurse" activity, the instructor led a large group discussion on Day 4 of the unit, just before

the group quiz. The discussion took about 18 minutes. To guide her discussion, she wrote the following two questions on the board:

(1) How are people/subjects selected to be in the study at all?

(2) How are subjects selected to be in the treatment group?

The instructor clarified that the second question was asked only *after* the sample had been taken. When a student asked a question about "random selection," the instructor clarified that "random selection" and "random sampling" referred to the same design.

The instructor asked students to discuss for three minutes what random sampling and random assignment allow researchers to say. Then, she revisited a previous quiz that students had completed, called "Dolphin Therapy." In this quiz, students had examined data from an experiment which used a volunteer sample of adults ages 18-65 with mild to moderate depression, and who were off of their medication. The experimenters had taken all subjects to the beach in Honduras and randomly assigned half of them to swim with dolphins and the other half to spend time on the beach without swimming with dolphins, as a control group. The instructor asked students to recall this example and mentioned that previously, she had accepted student answers that made claims such as "swimming with dolphins improves depression." But now, they needed to consider the study design in order to decide what claims were acceptable.

After allowing students time to discuss in small groups, the instructor brought students back to a large group discussion. When she asked about generalizations, students were quick to identify that one could not generalize to all humans and that obtaining a random sample of all humans would be impossible. A few students said that it was safe to

generalize to a more limited population of 18-65 year olds with mild to moderate depression taking no medication. One student brought up that it made a difference how they were recruited. The instructor agreed, pointing out that other variables such as participants' socioeconomic status or the region where participants were recruited could make these participants different from the general population of 18-65 year olds with mild to moderate depression taking no medication.

When the instructor asked about making causal claims, there was more hesitation than when students were discussing generalization. One student was skeptical of making causal claims, because she did not realize that the control group also went to the beach in Honduras, and thought that a beach vacation might be the variable influencing the results, rather than the dolphin therapy. After the instructor corrected this misinterpretation, and asked students whether they could make causal claims, most students nodded yes. One student correctly went on to explain that due to random assignment, the only difference between the groups was likely to be the treatment, rather than another variable. At the conclusion of this discussion, the instructor handed out the group quiz.

4.2.5 Results from classroom group quiz observation

While students took the quiz in the three in-class sections, the researcher and co-observer walked around to observe student groups. It was sometimes difficult to distinguish conversations due to so many students talking at once. Therefore, each observer tried to focus on one group at a time for a few minutes, wrote down observations of how students were discussing the questions, and wrote down approximately how long it took most groups to turn in their quiz. In all three sections, nearly all of the groups finished the quiz within half an hour of the quiz start time.

At the beginning of the quiz, a few students in section 2 were observed jotting notes down before looking at any of the questions. For example, some students drew the 2x2 table they remembered from the "Scope of Inferences" reading (Table 3.2) that clarified what conclusions could be made from studies that had random sampling, random assignment, both, or neither. A few students wrote down "random sampling -> generalization" and "random assignment -> causation."

Overall, many students were overheard giving correct answers to questions, such as pointing out that since random sampling of U.S. adults was used in a study, one could publish a stated headline making a generalization to U.S. adults. Students were also frequently overheard correctly stating that when a study did not use random assignment, headlines that made causal claims were not appropriate.

Question #3 on the quiz sparked much group discussion, compared to the other questions. Before reading this question, students were given a scenario in which 42 nutritionists at an ice-cream social were randomly assigned a small or large bowl, and the response variable of how much ice cream they served themselves was measured. Question #3 asked students whether confounding variables were likely to explain a significant difference in amount of ice cream served between those who had small bowls and those who had large bowls. Some students were initially observed talking about what other confounding variables could exist, such as diet and how much people liked ice cream. Later on, some students were seen erasing or crossing out answers, and discussing that the purpose of random assignment was to balance out confounding variables.

Some student groups were observed discussing their own knowledge about the context of the question, and using that to shape their answers. For example, question #5 on

the quiz asked students whether it was appropriate to state that a higher GPA would get students into medical school (i.e. making a causal claim). Some student groups discussed that medical school admissions involved more than just grades, such as essays, extracurricular activities, and other factors, without addressing the lack of random assignment. Other students were more critical of the headlines than the researcher anticipated, such as claiming that the headline "In U.S., Moderate Drinkers Have Edge in Emotional Health" from question #1 was inappropriate because the sample was only U.S. *adults*, and because being more likely to "experience positive emotions" was not the same thing as having better emotional health.

In general, groups did not appear to struggle a great amount in coming to consensus on answers. However, some students were observed disagreeing about how to interpret headlines. Some groups could not come to a consensus on whether a headline made a generalization and/or a causal claim. For example, students were often observed interpreting the previously mentioned headline "In U.S., Moderate Drinkers Have Edge in Emotional Health" as meaning that drinking *caused* people to have better emotional health. (The quiz clarified that this headline implied that "those who drink moderately tend to have better emotional health," but some students still interpreted "tend to" as causal language.) The subsequent question asked students whether it was appropriate to recommend that American adults consider drinking alcohol to increase positive emotions. Occasionally, after students read this second question, they realized that the two questions they were looking at in this context were different, and corrected their response to the first question. Students also were observed struggling to come to a consensus on interpreting the headlines presented in questions #5 and #6 about medical school admissions. Some groups were

observed debating whether generalizations or causal claims could be made by each headline. Results from analysis of the group quiz responses are discussed in section 4.4.

Survey Incentives **activity**

The "Survey Incentives" activity involved a context in which random sampling and random assignment are both possible. This activity was similar to the "Sampling Countries" and "Strength Shoe" activities in that students carried out random sampling and random assignment for many trials. However, they did random sampling and random assignment within the same context, and were asked to compare the two types of study designs and conclusions at the end. This was the last activity of the unit. For the in-class sections, this activity happened on the class period following the group quiz, and was the last day of the unit. Online, the activity happened the last week of the unit, and the same week as the group quiz.

Survey Incentives: Classroom Observation Checklist

Table 4.7 shows an abbreviated observation checklist for "Survey Incentives" with the number of times each element was checked. The full lesson plan checklist for this activity is in Appendix E4. Since "Survey Incentives" happened on the last day of the unit and could not carry over to the next class period, four wrap-up questions were designated essential "key wrap-up questions," bolded in the table below. There were also some encouraged wrap-up questions, shown in italics in the table, which were considered important but not essential. In all sections, these key wrap-up questions were discussed, although the instructor of section 2 did not explicitly address how randomness was different in random assignment versus random sampling. However, she did discuss questions about

the differences between random sampling and random assignment, and what conclusions one can make from each.

Table 4.7
Summary of observation checklist results for "Survey Incentives" activity.

Element from lesson plan[a]		Section 1	Section 2	Section 3
L1.	Instructor briefly introduces the activity[b]	2	2	2
L2.	Instructor asks students to turn off animation	2	2	2
L3.	Instructor asks students to work through activity in groups	0	2	2
S1	Students say "randomly sample" with no detail	0	1	0
S1A.	How can the mayor take a random sample from her list?	0	1	0
S1B.	What steps would you advise her to take?	0	1	0
S2.	Students say they will use *TinkerPlots*™ to sample randomly (no detail on how)	0	1	1
S2A.	Instructor asks them to describe how to set up sampler to randomly sample	0	2	0
S3.	Students ask what "similar" means when comparing sample to population	0	1	0
S3A.	Do population and sample look similar?	0	0	0
S3B.	Do you expect sample will be similar to population?	0	0	0
S4.	Students struggle with question on whether random sampling appears unbiased	0	0	0
S5.	Students struggle to pick a confounding variable to explore	1	1	0
S6.	Students struggle to explain why confounding variable would affect results	0	0	1
S6A.	Which variable do you think would affect how people respond?	0	0	0
S6B.	How do you think [age, income, hours worked] might influence willingness to respond?	0	0	0
S7.	Students say random assignment is not possible with uneven sample size in each group	1	1	0
S7A.	How can you make group sample sizes as even as possible?	0	0	0
S8.	Students just say "randomly assign" with no detail	1	0	1
S8A.	What detailed steps would you take to randomly assign?	1	0	1
S9.	Students say they will use *TinkerPlots*™ to randomly assign (no detail on how)	0	0	0
S10.	Students ask what "similar" means when comparing group means.	1	0	2

Element from lesson plan[a]	Section 1	Section 2	Section 3
S10A. Do you expect the two groups will have similar characteristics?	0	0	1
S11. Students have trouble answering whether random assignment is effective for balancing out confounding	0	0	0
S12. Students say random assignment is not effective for balancing out confounding variables	0	0	0
S13. Students struggle to summarize difference between random sampling and random assignment	0	0	0
S14. Students cannot differentiate between random sampling and random assignment	0	0	0
L4. What variable did you choose to collect statistics for in question #10?	0	0	0
L5. Where was your plot of sample statistics centered?	0	0	0
L6. Why did you expect it to be centered at this value?	0	0	0
L7. What does it mean for a sampling method to be unbiased?	0	0	0
L8. Why does an unbiased sampling method allow us to generalize?	0	0	0
L9. Why is random sampling better than dropping surveys in mailboxes?	1	0	0
L10. What is the treatment variable?	0	0	0
L11. What is the response variable?	0	0	0
L12. What confounding variable did you explore?	0	0	0
L13. Where was plot of differences in means centered?	0	0	0
L14. Why does it make sense plot was centered at 0?	0	0	0
L15. What is the purpose of using random assignment in this study?	0	0	0
L16. What is the difference between random assignment and random sampling?	2	2	2
L17. How is the randomness different in each case?	2	0	2
L18. Why does random sampling allow us to generalize to the population?	2	2	2
L19. Why does random assignment allow us to make causal claims?	2	2	2

[a] Elements in bold indicate essential parts of the lesson plan (key questions). Elements in italic indicate recommended, but not required, wrap-up questions. Elements not bolded indicate suggested components or suggestions for potential issues that might arise and how to address them.
[b] Elements are numbered. Those that begin with *L* indicate large group discussion components of the lesson, where those that begin with *S* indicate small group discussion suggestions.

Survey Incentives: In-class instructor implementation

Each in-class instructor began the class period differently. The instructor of sections 1 and 3 had encountered a study reported in the media about red meat causing cancer. She

used this real-world example to lead a discussion about considering study design and conclusions. At the beginning of each of her classes, the instructor of sections 1 and 3 showed a video from a popular news channel reporting the results of a study that concluded that red meat could shorten one's lifespan. The reporters recommended that people replace red and processed meat in their diet with chicken, fish, and other proteins. The instructor of section 2 did not use this study to prompt large group discussion, because in her pre-quiz discussion she had already referred to a real study about dolphin therapy for patients with depression.

After showing the news clip, the instructor asked students in section 1 to first discuss in small groups what they thought about the study. In section 3, after the video, the instructor led a large group discussion about the study without first giving students time to discuss in small groups. In both sections, students brought up in large group discussion that the lack of random sampling meant that the subjects were not necessarily representative of the U.S. population. In both sections, students also mentioned the lack of ability to make causal claims due to confounding variables. In order to test for a possible misconception, the instructor asked in section 1: "If the study had been a random sample, could I have cause and effect?" There were mixed responses, but most students shook their heads "no" and one student clarified that a random sample allows one to generalize, not make causal claims. In both sections 1 and 3, the instructor concluded this short discussion of the article by mentioning that the media often over-generalizes results from studies and makes ungrounded causal claims.

Because of the extra video and discussion, sections 1 and 3 started the activity about 20 minutes into the class period, whereas section 2 started the activity just a few minutes

into the class period after the instructor made some brief announcements. As requested in the lesson plan, the two instructors briefly introduced the activity by mentioning that the activity would be wrapping up ideas of random sampling and random assignment. They pointed students to the necessary *TinkerPlots*™ files to download and reminded them to turn off the software animation to save time.

When instructors saw that groups finished the activity early, they had discussions with those groups and asked them about their responses to the last few questions in the activity (those that asked about the differences between random sampling and random assignment). Also, the instructor of sections 1 and 3 often asked groups: "If the mayor does random sampling, can she make a causal claim?" In general, students quickly answered "no" to this question and mentioned that random assignment was necessary. Although overall, students correctly said that random sampling allows the mayor to generalize and random assignment allows her to make causal claims, few went into details about why these things were true. However, when the instructor prompted students to explain why each study design led to each type of conclusion, many of them explained that a random sample would be representative of the population and random assignment helped to balance out confounding variables. Although the activity asked students to write a "short report" at the end about the differences between random sampling and random assignment, most students were observed only writing a couple of sentences. The instructor of sections 1 and 3 saw this, and asked students to write a report and e-mail it to her by the end of the class day. The instructor of section 2 did not do this.

In all three in-class sections, wrap-up started about 15-17 minutes before class ended. In all three sections, the instructors asked their classes to summarize the main ideas

of the activity. Students in each section gave similar answers, saying that the purpose of the activity was to distinguish between random sampling, which is necessary for generalization, and random assignment, which is necessary for causation. Also in all sections, the instructors skipped most of the suggested wrap-up questions and instead focused on the main ideas of the four key questions at the end of the lesson plan. They also addressed questions that had come up frequently for students during the activity.

Since the instructors had seen that students had not given much detail on how to take a random sample or conduct a random assignment, they asked students to do this in large group discussion. With some prompting, students gave correct answers involving mechanisms such as drawing names out of a hat, or using a computer to select random numbers. In section 2, many students had talked about making sure all groups were represented. The instructor of section 2 spoke briefly about stratified sampling. She said that although stratified random sampling was recommended when it was very important to have representation of certain groups, this would give less flexibility for the rest of the random selection, and some of the "power" of the randomness would be lost.

In all sections, the in-class instructors asked students why random sampling allows for generalization and random assignment allows for causation. Some students in each in-class section brought up the fact that since each person is equally likely to be in the sample, the sample was not biased such that people of certain groups were more likely to be in the sample than others. For random assignment, students also explained correctly that random assignment helps to balance out confounding variables, so that they do not influence the response variable and we can attribute differences to the treatment variable.

Additional questions from students came up during large group discussion. In section 1, a student asked whether a random sample could end up being all females, just by chance. The instructor revisited the idea she had presented during "Sampling Countries" discussion that most of the time, random samples were representative and sample statistics were near the parameter, but rarely, unusual samples happen. A student in section 2 asked the instructor to clarify the difference between "association" and "causation," and the instructor led a short discussion about this. A student brought up ethical issues in section 3, and the instructor gave an example of how studies involving smoking cannot ethically involve random assignment, but associations found from these studies can still be useful.

Table 4.8 summarizes the similarities and differences seen in the three sections of the class during the "Survey Incentives" activity and discussion.

Table 4.8
Summary of methods used during large-group discussion of "Survey Incentives" activity

Method	Section 1	Section 2	Section 3
Began class with showing a video on red meat and cancer	X		X
Allowed time for small-group discussions on study about red meat and cancer	X		
Led large-group discussion on study about red meat and cancer	X		X
Led large-group discussion on study about dolphin therapy and depression (previous class day)		X	
Began activity about 20 minutes into class period	X		X
Began activity less than 5 minutes into class period		X	
Introduced activity by mentioning that it would wrap up ideas of random sampling and random assignment	X	X	X
Had discussions about the activity with groups who finished early	X	X	X
Asked groups who finished early: "If the mayor does random sampling, can she make a causal claim?"	X		X
Asked students to write a short report and e-mail it to her	X		X
Began wrap-up 15-17 minutes before end of class	X	X	X
Skipped most suggested wrap-up questions, focusing only on key questions	X	X	X
Asked students for details about how to randomly sample and how to randomly assign	X	X	X

Method	Section 1	Section 2	Section 3
Asked students about why each study design allows for each type of conclusion	X	X	X
Briefly discussed stratified random sampling		X	
Addressed the idea of unusual random samples	X		
Led brief discussion on the difference between "association" and "causation"		X	
Addressed ethical issues of experiments			X

Note. An "X" in a cell indicates the instructor of that section used the corresponding method

Survey Incentives: Online instructor implementation

The online instructor monitored the discussion of the "Survey Incentives" activity, addressing issues that came up as needed. For example, when students suggested non-random methods of sampling such as taking every 10^{th} name, the instructor challenged the group of students to think about whether this method was truly random. The last two sets of group discussion questions online involved discussing the differences between random sampling and random assignment in context, and why they allowed for generalization to the town population and causal claims about the survey incentive, respectively. Since some students made mistakes when talking about these concepts (such as using the words "random sampling" when random assignment was the pertinent design), the instructor posted a clarification for many groups, distinguishing between the random sampling from the population that happened at the beginning, and the random assignment to groups that happened after the sample was selected.

The online instructor wrote a brief wrap-up of about 170 words, first addressing a few main points. He addressed the misconception that one can never generalize to a population with a small sample, and reminded students of the "Sampling Countries" activity where they learned that a small random sample was better than a large biased sample. He also emphasized the idea that in real studies, when a random sample is taken

there is only one random sample. Also, when an experiment is conducted, only one random assignment is performed. Therefore, researchers need to be able to trust that their method of sampling will tend to produce an unbiased estimate, and their method of random assignment will tend to balance out confounding variables. After addressing these points, the instructor shared a student's exemplary answer, after having obtained the student's permission. This answer (of about 360 words) clearly described why random sampling helps the mayor to generalize, why random assignment helps the mayor to make causal claims, and how these two concepts are different.

Survey Incentives: Observations of activity discussions

As seen in Table 4.7 above, most of the issues anticipated by the researcher arose during the activity in at least one class section. The instructors noticed that students were not giving enough detail about how to take a random sample or conduct a random assignment, so they addressed this during large-group discussion. Online students were also not specific on how to carry out each method. When asked about how to take a random sample, some students suggested non-random ways of sampling. For example, some suggested hand-picking participants from different neighborhoods to make sure that people of different groups were represented, or taking every *n*th name from a list (systematic sample).

Students were asked in the activity whether they thought their sample looked similar to the population, and whether the two randomly assigned groups looked similar with respect to a confounding variable. As anticipated, some students in sections 1 and 2 asked the instructors what "similar" meant, and the instructors responded by asking them to look at "the whole variable" (possibly meaning to look at the entire distribution of the

variable for the sample and for the population, rather than just comparing the means) and judge this for themselves.

After the sampling portion, the activity asked students to choose a confounding variable and explain why it could be a confounding variable. Many students in all sections were observed choosing income, but some appeared unsure about why this would be a confounding variable. Others were observed correctly explaining that the $20 incentive would be more appealing to those of lower incomes than to those of higher incomes.

Students were asked to randomly assign 25 subjects to groups. The odd sample size was chosen on purpose to target the possible misconception that sample sizes needed to be equal. Students in sections 1 and 2 were observed questioning whether they could randomly assign 25 subjects, because the sample sizes were unequal. In class, the instructor or TA clarified for some students that real studies are often not balanced, but that is all right because averages are being compared. Online, when faced with the unequal sample sizes, some students suggested nonsensical ways of dividing students into groups, such as creating 5 groups of 5 subjects each for the random assignment. However, most online students correctly suggested assigning 12 subjects to one group and 13 subjects to another.

Although the main point of the activity was to help students distinguish between random sampling and random assignment, a few students, especially online, still showed some confusion between the two. For example, online, various students used the term "random sampling" when they should have said "random assignment," such as saying that the "random sampling" tended to balance out confounding variables between the two groups. In class, students sometimes spoke of the importance of randomly assigning participants in the sample to groups during the sampling part of the activity, but did so

correctly. Online, a noticeable amount of students gave answers that still showed confusion between random sampling and random assignment. For example, two online students' answers claimed that random assignment was necessary for generalizing and random sampling was necessary for making causal claims. Two others said that random assignment allows us to both generalize and make causal claims, and one said that random sampling allows for both generalization and causation. Although in class it was not possible to hear every group's conversation, incidents of students making these incorrect statements were not common in the observer notes.

Most students in class did not write much for their "short report" about the distinction between random sampling and random assignment, except for students in sections 1 and 3 who were asked to e-mail their reports to the instructor. Online, however, students wrote more. Most online students correctly wrote that random sampling was needed for generalization and random assignment was need for making causal claims. However, only about one-third of online students had clear explanations as to *why* this was the case, such as discussing the need for a representative sample and the need to balance out confounding variables.

In summary, the nine issues below arose during the activity. The first six were addressed in the lesson plan. Of the last three, the last two show confusion between random sampling and random assignment, a problem that was also anticipated to happen during the unit, though not necessarily anticipated to be predominant during the last activity of the unit.

- Not describing in detail how to take a random sample
- Not describing in detail how to conduct a random assignment

- Difficulty judging whether the distribution of a variable for a sample is similar to that of the population
- Difficulty judging whether the distribution of a variable is similar between two groups
- Believing that random assignment cannot be done with two groups of unequal sizes
- Not fully explaining how a specific variable might be a confounding variable
- Suggesting non-random methods of sampling, such as purposefully picking people to ensure groups are represented
- Suggesting that with random assignment, one can both generalize *and* make causal claims
- Confusing the terms "random sampling" and "random assignment," such as suggesting that random sampling balances out confounding variables

Despite these issues, confusion between random sampling and random assignment appeared to be less prevalent in this activity than in previous activities. Although the online class appeared to have more problems than the in-class sections with distinguishing between random sampling and random assignment, most students appeared to be able to explain that random sampling helped with generalizing to the town population and random assignment helped with enabling causal claims about the survey incentive.

4.3 Results from the Inferences from Design Assessment

The IDEA was administered to students as a pretest just before the study design unit began, and as a posttest upon the conclusion of the unit. This section will describe

results from quantitative analysis of the data from the administration of IDEA, including reliability analyses, examination of scores, and examination of individual items. IDEA contained 22 total items, 9 of which were related to concepts of random sampling and generalization (which will be referred to as the sampling items), and 13 of which were related to concepts of random assignment and causation (which will be referred to as the assignment items). The total score (number correct out of 22) was computed for each respondent. Also, the score (number correct out of 9) from the sampling items (sampling subscore) and the score (number correct out of 13) from the assignment items (assignment subscore) were computed for each respondent.

4.3.1 Reliability

The reliability of an assessment refers to the consistency of measurements taken from an individual (AERA, APA, & NCME, 1999; Thorndike & Thorndike-Christ, 2010). In other words, reliability is the fraction of total test score variance that is true score variance, rather than variance due to error. After consultation with a measurement expert at the University of Minnesota, it was decided to compute coefficient omega (McDonald, 1999) to measure reliability. Omega hierarchical (ω_h), based upon the sum of squared loadings on one general factor, represents the reliability within which the test measures a single construct (Revelle & Zinbarg, 2009). Omega total (ω_t), based upon the sum of squared loadings on all the factors, represents the proportion of test variance due to all common factors (Revelle & Zinbarg, 2009). Both omega coefficients are reported in Table 4.9 below for the IDEA pretest and posttest, for both the Sampling and Assignment subscales, and for the total score.

Both omega hierarchical (the reliability within which the set of items measures a single factor) and omega total (the proportion of test variance due to all common factors) can be reported as reliability coefficients (Revelle & Zinbarg, 2009). According to Nunnally (1978, p. 245), a reliability coefficient of .70 may be adequate in the early stages of research, although in basic research it is preferable to have a reliability of .80 or greater. Overall, the omega hierarchical coefficients indicate low reliability in measuring a single factor for the IDEA test as both a pretest and posttest, both as a whole and for each subscale. The omega total coefficients also indicate low reliability of the IDEA test when used as a pretest, both as a whole and for each subscale. The omega total coefficient for the IDEA posttest as a whole approaches Nunnally's (1978, p. 245) suggestion of .80 or greater for basic research, but for each subscale on the posttest, the omega total coefficient indicates modest reliability.

Table 4.9
Values of Omega for IDEA pretest and posttest, and for Sampling and Assignment subscales

		Sampling subscale	Assignment subscale	Total score
Pretest	Omega hierarchical (ω_h)	0.30	0.24	0.26
	Omega total (ω_t)	0.56	0.57	0.63
Posttest	Omega hierarchical (ω_h)	0.36	0.36	0.46
	Omega total (ω_t)	0.72	0.68	0.79

4.3.2 Examining correlations between subscale scores

In order to decide whether the sampling and assignment subscales contribute unique information about students' scores, the correlation between these two subscales was examined as advised by consultation with a measurement expert at the University of

Minnesota. The Pearson correlation between the two subscores was 0.41 for the pretest and 0.79 for the posttest.

However, in order to account for measurement error, each correlation was corrected for attenuation due to measurement error (Spearman, 1904; as cited in Charles, 2005). The corrected correlation r_{xyc} was computed by using this equation:

$$r_{xyc} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}} \quad (4.1)$$

Where r_{xy} is the observed Pearson correlation between the two subscores, r_{xx} is an estimate of the reliability of the sampling subscores, and r_{yy} is an estimate of the reliability of the assignment subscores. A very high correlation would indicate that the two scores do not contribute unique information, and would suggest examination of only the total score.

Using the omega hierarchical values as reliability estimates, for the pretest, the corrected correlation for attenuation was computed:

$$r_{xyc\_pretest} = \frac{0.41}{\sqrt{0.30}\sqrt{0.24}} = 1.53 \quad (4.2)$$

For the posttest, this calculation is:

$$r_{xyc\_posttest} = \frac{0.55}{\sqrt{0.36}\sqrt{0.36}} = 1.53 \quad (4.3)$$

Both of these correlations were above 1.0, outside the possible range for a correlation. However, Charles (2005) denotes that correlation corrections for attenuation above 1 can occur under certain conditions, such as when reliability is underestimated.

When omega total was instead used as an estimate of reliability, the corrected correlation for attenuation for the pretest was calculated:

$$r_{xyc\_pretest} = \frac{0.41}{\sqrt{0.56}\sqrt{0.57}} = 0.73 \quad (4.4)$$

For the posttest, this calculation is:

$$r_{xyc\_posttest} = \frac{0.55}{\sqrt{0.72}\sqrt{0.68}} = 0.79 \quad (4.5)$$

As previously stated, both omega hierarchical and omega total can be reported as reliability coefficients (Revelle & Zinbarg, 2009). Therefore, the most conservative estimate for the corrected correlation was taken to make a decision about whether or not to consider both subscores in analysis. When omega total was used, the sampling and assignment subscore correlations, corrected for attenuation, appeared to be moderately high, but not large enough to say that the two subscales were very highly correlated and thus might not contribute unique information. It appears that it is worth examining the sampling and assignment subscores separately, in addition to examining the total score as well.

4.3.3 Descriptive analysis of IDEA test scores

Descriptive statistics were computed for IDEA pretest and posttest scores, for the assessment as a whole and also for the sampling and assignment subscale scores separately. Table 4.10 below displays descriptive statistics for all students who took the assessment. The average and median scores for the total score and each of the two subscores increased from pretest to posttest. The standard deviation on pretest and posttest for each of the two subscales was similar. The standard deviation for the total score was slightly larger for the pretest than the posttest.

Table 4.10
Descriptive statistics of IDEA pretest and posttest scores

		Mean	SD	Min.	1st quartile	Median	3rd quartile	Max.
Pretest (n = 131)	Total score (out of 22)	14.55	2.79	7	13	15	16.5	22
	Sampling subscore (out of 9)	4.83	1.62	1	4	5	6	9
	Assignment subscore (out of 13)	9.72	1.71	3	9	10	11	13
Posttest (n = 130)	Total score (out of 22)	17.88	3.05	7	17	19	20	22
	Sampling subscore (out of 9)	6.62	1.7	0	6	7	8	9
	Assignment subscore (out of 13)	11.26	1.75	5	11	12	12.75	13

4.3.4 Comparing the four sections on their IDEA performance

The four sections of the course were compared on their IDEA pretest and posttest performance. Table 4.11 below summarizes the means and standard deviations for each section's scores on the pretest and posttest.

Table 4.11
Means and standard deviations of IDEA scores divided by section

		Section 1 (n = 39)		Section 2 (n = 32)		Section 3 (n = 24)		Section 4 (n = 36)	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
Pretest	Total score (out of 22)	15.41	2.81	14.56	2.34	14.88	3.03	13.39	2.69
	Sampling subscore (out of 9)	5.18	1.71	4.78	1.72	5.21	1.32	4.25	1.48
	Assignment subscore (out of 13)	10.23	1.61	9.78	1.43	9.67	2.01	9.14	1.69
		Section 1 (n = 39)		Section 2 (n = 30)		Section 3 (n = 28)		Section 4 (n = 33)	
Posttest	Total score (out of 22)	18.13	3.08	17.57	2.92	18.21	2.86	17.58	3.35
	Sampling subscore (out of 9)	6.59	1.67	6.53	1.61	6.89	1.69	6.48	1.89
	Assignment subscore (out of 13)	11.54	1.82	11.03	1.79	11.32	1.61	11.09	1.79

According to the descriptive statistics broken down by section, all four sections appear similar to each other in terms of how they scored for the total score and each of the two subscales, at each time point. On the pretest, the online section appeared to have the lowest mean total score, sampling subscore, and assignment subscore. However, on the posttest, the four sections appear to be more similar to each other.

In order to test whether significant differences existed between sections, one-way ANOVA analyses were conducted. Since the ANOVA analyses were done for the pretest, for the posttest, and for the total score, a Bonferroni adjustment was used. The familywise Type I error rate was set to be alpha = 0.05, and thus each of the three tests was conducted at alpha = 0.017. For the pretest, the analyses revealed that there were significant

differences at α = .017 among the sections for the total score ($F = 3.63$, $p = .002$), but not for the sampling score ($F = 2.70$, $p = .049$), or for the assignment score ($F = 2.68$, $p = .050$).

Since the ANOVA analysis revealed significant differences between sections for the total score on the pretest, pairwise t-tests were then conducted to explore which sections were different from each other on the pretest total score, using Bonferroni multiple comparisons adjustments. The Bonferroni-adjusted *p*-values for each pair of sections is shown in Table 4.12 below. Only sections 1 and 4 were significantly different from each other (Bonferroni-adjusted $p = .010$).

Table 4.12
Bonferroni-adjusted p-values for pairwise comparisons of total IDEA pretest score

Section	1	2	3
2	1.00	--	--
3	1.00	1.00	--
4	.010	.462	.237

For the posttest, one-way ANOVA analyses revealed that there were no statistically significant differences at any reasonable significance level among the sections for the total score ($F = 0.41$, $p = .747$), for the sampling subscore ($F = .60$, $p = .615$), and for the assignment subscore ($F = .33$, $p = .802$).

4.3.5 Pretest to posttest changes in IDEA test scores

Changes in scores from pretest to posttest were examined for the total score, sampling score, and assignment score. The pretest score was subtracted from the posttest score to find the difference in scores for each student. Descriptive statistics of these differences

were computed for all 125 students who completed the IDEA pretest and posttest, and are presented in Table 4.13 below.

Table 4.13
Descriptive statistics of IDEA differences (posttest – pretest) for n = *125 students*

	Mean	SD	Min.	1st quartile	Median	3rd quartile	Max.
Difference in total score (out of 22)	3.30	2.94	-8	2	3	5	14
Difference in sampling subscore (out of 9)	1.75	1.79	-4	1	2	3	6
Difference in assignment subscore (out of 13)	1.55	1.87	-5	1	2	3	10

On average, students increased their score from pretest to posttest. More than 75% of students increased their total score, as well as their sampling and assignment subscores.

In order to test whether the increases from pretest to posttest were statistically significant, a paired *t*-test was conducted for each score. Additionally, Cohen's *d* values were computed to examine the effect size of the difference for each score. Mean differences, *p*-values, confidence intervals for the mean differences, and Cohen's *d* values are presented in Table 4.14 below.

Table 4.14
Results from paired t-tests of IDEA differences (posttest – pretest) for n = 125 students.

	Mean Difference	SD	*t*	*p*	Cohen's *d*
Difference in total score (out of 22)	3.30	2.94	12.57	<.001	1.12
Difference in sampling subscore (out of 9)	1.75	1.79	10.97	<.001	0.98
Difference in assignment subscore (out of 13)	1.55	1.87	9.29	<.001	0.83

Since three paired *t*-tests were conducted, the family-wise Type I error rate was set at alpha = .05, making the alpha level for each test .017. The total score and each subscore increased significantly at $\alpha = .017$ from pretest to posttest. The effect size indicates that for the total score, the average score increased by just over 1 standard deviation, and for the sampling subscore, the average score increased by just under 1 standard deviation. For the assignment subscore, the increase was slightly lower, at 0.83 standard deviations.

4.3.5.1 Comparing the four subsections on their changes from pretest to posttest

The four different sections of the class were compared on their change in scores from pretest to posttest for total score, sampling subscore, and assignment subscore. First, means and standard deviations were computed to descriptively examine how the four sections of students were similar or different in their average differences and in the variability of those differences. These descriptive statistics broken down by section are displayed in Table 4.15 below.

Table 4.15
Means and standard deviations of IDEA differences in scores (pretest-posttest), by section

	Section 1 (n = 38)		Section 2 (n = 30)		Section 3 (n = 24)		Section 4 (n = 33)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Difference in total score (out of 22)	2.50	2.61	2.90	2.92	3.62	2.06	4.36	3.56
Difference in sampling subscore (out of 9)	1.29	1.56	1.67	1.71	1.75	1.29	2.36	2.25
Difference in assignment subscore (out of 13)	1.21	1.83	1.23	1.81	1.88	1.39	2.00	2.18

For all three scores, the students in the online section (section 4) appeared to have slightly higher average gains than the students in the other three sections.

In order to test whether the differences in mean change significantly differed by section, a one-way ANOVA was conducted for each set of score differences (sampling subscore, assignment subscore, and total score). Since three tests were conducted, the familywise Type-1 error rate was set at α =.05, so the alpha level for each test was adjusted to .017. For the differences in total score, the ANOVA did not reveal a significant difference at α = .017 among the sections (F = 2.78, p = 0.044). Also, one-way ANOVAs did not reveal significant differences between sections in their changes from pretest to posttest for the sampling subscore (F = 2.23, p = .088) or for the assignment subscore (F = 1.61, p = .191).

4.3.6 Pretest to posttest changes in IDEA individual items

The next step in examining changes from pretest to posttest was to explore changes in correct responses for individual items. Responses to each item on the IDEA pretest and posttest were coded as 0 to indicate an incorrect answer, and 1 to indicate a correct answer.

Then, four different categories of response patterns were identified. The first category, labeled "incorrect," represents answers that were incorrect on both pretest and posttest. The second category, labeled "decrease," represents answers that were correct on the pretest but incorrect on the posttest. The third category, labeled "increase," represents answers that were incorrect on the pretest but correct on the posttest. The fourth category, labeled "pre & post," represents answers that were correct on both pretest and posttest.

In addition, as the data consist of two dependent samples (students' responses at two time periods, pretest and posttest), McNemar's test was used to examine whether the change from pretest to posttest for each item was statistically significant. Because some items contained very low percentages of students answering incorrectly, the chi-square approximation may not hold; therefore, an exact McNemar's test was used to test significance for each item. A family-wise Type I error rate was set at $\alpha = 0.05$ across the 22 McNemar's tests conducted, and using the Bonferroni method, a per test Type I error limit was set at $\alpha_c = .002$. The full table of percentages of students who fell into each response pattern category, along with *p*-values for each item, are presented in Appendix K.

Items were classified into one of three categories: (1) Items with high percentages of students with correct answers on both pretest and posttest, (2) items with statistically significant increases in percentage of students with correct responses from pretest to posttest, and (3) items with non-statistically significant increases in percentages of students with correct responses from pretest to posttest. There were no items with statistically significant decreases from pretest to posttest.

4.3.6.1 Items with high percentages of students with correct answers on both pretest and posttest

For nine items, over 80% of students provided correct answers both before the study design curriculum began, and after the curriculum was over. The percent of correct responses for each of these items, along with their *p*-values from the McNemar's tests, are shown in Table 4.16 below. Two of these items (1 and 7) were in the sampling section of the IDEA assessment, while the other seven items (10, 11, 12, 13, 14, 15, and 20) were on the assignment section. Seven items showed an increase in performance from pretest to posttest, but the increase was not statistically significant at α_c = .002 for any of them. Two items (10 and 14) showed a slight decrease in performance from pretest to posttest, but the decrease for each item was less than 5 percentage points and did not approach statistical significance.

Table 4.16
Items with 80% or more students correct on the pretest and the posttest

Item	Measured Learning Outcome	*n*	% of Students Correct		McNemar's test *p*
			Pretest	Posttest	
1	(One of two-item set): Ability to identify the sample.	125	90.4	92.0	.815
7	Ability to understand that random sampling is preferable to non-random methods of sampling for a sample to be representative of the population.	125	88.8	96.8	.006
10	Ability to determine what type of study was conducted (observational or experimental).	125	94.4	90.4	.302
11	Ability to understand that a randomized experiment is needed to answer research questions about causation.	125	94.4	96.8	.375

12	(Four-item set): Ability to	125	92.8	96.0	.387
13	distinguish between	125	90.4	92.0	.804
14	statements that make causal	125	88.8	86.4	.664
15	claims and statements that make association-only claims	125	94.4	96.8	.549
20	(One of three-item set): Ability to understand that random assignment is the best way to balance out groups with respect to confounding variables.	124	88.7	91.9	.541

More than 90% of students correctly identified a sample from a study description on both the pretest and the posttest (item 1). Also, on both pretest and posttest, more than 85% of students correctly recognized that a non-random sample was likely biased (item 7). Item 7 showed an increase in performance of eight percentage points from pretest to posttest, but the increase was not statistically significant (after adjusting for multiple comparisons).

The remaining seven items with high percentages of correct answers were in the assignment section of the assessment. On both pretest and posttest, over 90% of students correctly recognized an experimental study (item 10) and indicated that in order to establish causation, a study design using random assignment is preferred over observational studies (item 11). In general, students demonstrated a high capacity for being able to distinguish between statements that make association-only claims from statements that make causation claims (items 12-15). Also, on both pretest and posttest, the great majority of students correctly recognized that using random number sequences to assign students to treatments was a valid method of random assignment (item 20).

4.3.6.2 Items with statistically significant increases in percentage of students with correct responses from pretest to posttest

There were nine items with an increase in student performance that was statistically significant after adjusting for multiple comparisons. The percent of correct responses for each of these items, along with their *p*-values from the McNemar's tests, are shown in Table 4.17 below. The first five items in Table 4.17 (items 2, 3, 4, 5, and 6) were from the sampling section, and the remaining four items (16, 18, 21, and 22) were from the assignment section.

Table 4.17
Items with a statistically significant gain from pretest to posttest

Item	Measured Learning Outcome	*n*	% of Students Correct		McNemar's test *p*
			Pretest	Posttest	
2	(One of two-item set): Ability to identify the sample and the population to which inferences can be made.	125	40.8	65.6	<.0001
3	Ability to understand what it means to make an appropriate generalization to a population, using sample data.	125	23.2	63.2	<.0001
4	Ability to understand the factors that allow (or do not allow) a sample of data to be representative of the population.	125	8.0	32.0	<.0001
5	Ability to understand when sample estimates may be biased due to lack of a representative sample.	125	70.4	86.4	.0005
6	Ability to understand that a small random sample is preferable to a larger, biased sample.	125	46.4	85.6	<.0001
16	Ability to understand that correlation does not imply causation.	125	28.0	77.6	<.0001
18	Ability to understand the purpose of random	125	32.0	77.6	<.0001

	assignment in an experiment: To make groups comparable with respect to all other confounding variables.				
21	(One of three-item set): Ability to understand that random assignment is the best way to balance out groups with respect to confounding variables.	122	60.7	79.5	.0006
22	Ability to recognize when a randomized experiment is the most salient research design for a particular research question.	124	79.8	91.9	.0015

Only one item on the IDEA instrument had fewer than 60% of students answering correctly on the posttest. This was item 4, which involved identifying a factor that does not allow for generalization of survey results. Fewer than 10% of students on the pretest correctly identified that the sample size of 500 was not a problem for generalizability. This percentage of correct answers only increased to almost 34% on the posttest. A high number of students, both on the pretest and posttest, indicated that all answer choices (sample size of 500, limited sampling frame, and low response rate) were problematic for making generalizations (see Appendix J).

There were four items that showed noticeably large gains from pretest to posttest. The gains for items 3, 6, 16, and 18 were all between 39 and 50 percentage points. The item with the largest gain, item 16, involved being able to recognize whether or not a causal claim can be made if a strong, statistically significant correlation is found in a study that is observational, and choosing the correct reason for this. On the pretest, just over one-quarter of students correctly identified that a causal claim cannot be made due to the lack of random

assignment. On the pretest, more than one-third of students instead indicated that the sample size was too small to infer causation, and more than one quarter indicated that random sampling allowed for causal claims to be made (see Appendix J). On the posttest, the most common error was indicating that random sampling allows for causal claims, with just over 10% of students choosing this option.

The item with the second largest gain was item 18, which involved identifying the purpose of random assignment to treatments. On the pretest, just over 30% of students answered the item correctly. The most popular answer choice on the pretest, with almost 40% of students choosing this option, was that random assignment would ensure that study participants are likely to be representative of the population (thus showing potential confusion between random sampling and random assignment). More than one-quarter of the students also incorrectly indicated on the pretest that random assignment would ensure that subjects were not likely to know whether they were getting the placebo. On the posttest, more than three-quarters of students correctly identified the purpose of random assignment as ensuring that all groups were likely to be similar in all respects except for the treatment variable. Still, almost 15% of students incorrectly indicated on the posttest that the purpose of random assignment was to ensure that study participants would likely be representative of the population.

Another item with a large gain was item 3, with an increase in correct answers of 40 percentage points from pretest to posttest. This item required students to recognize a statement that made a generalization to an appropriate population of interest. On the pretest, fewer than one-quarter of students correctly identified that using a random sample of students from a certain high school would allow for generalizations to be made to students

at that high school. Also on the pretest, just over half of students incorrectly indicated that one could only generalize to the sample that was taken, and just over one-quarter of students incorrectly indicated that it was appropriate to generalize to all high school students (see Appendix J). On the posttest, over 60% of students answered the item correctly, although just over one-quarter of them still indicated that it was only appropriate to generalize to the sample.

Item 6 also had a gain of almost 40 percentage points from pretest to posttest. This item involved recognizing that a study using a random sample (with a 100% response rate) was preferable for providing unbiased estimates than a study where the entire sampling frame was contacted and a higher number of responses was obtained, but with a low response rate. Fewer than 50% of students correctly identified that the study with the random sample was preferable to the study that contacted all of the sampling frame on the pretest, but this increased to more than 85% on the posttest.

Item 5, another item related to sampling and bias, had significant gains from pretest to posttest. Although just over 70% of students correctly identified on the pretest that a generalization statement was invalid due to lack of a representative sample, over 15% of them incorrectly identified the small sample size (10,000 out of a population of 500,000) as the reason why the statement was invalid (see Appendix J). On the posttest, the percent of correct answers increased to 86%, and only about 3% of students indicated the statement as invalid due to the "small" sample size.

It is noteworthy that a high percentage of students correctly identified a sample from a study on both pretest and posttest (item 1), but they had more difficulty identifying a population of interest (item 2). Only about 40% of students were able to identify the

population correctly on the pretest, with the remaining students either incorrectly identifying the statistic as the population, or the sample as the population. On the posttest, the percent correct increased to more than 65%, but 20% of students still incorrectly identified the sample as the population (see Appendix J).

Items 21 and 22 about randomized experiments also had significant gains, although for each of these items, over 60% of students answered the item correctly on the pretest. Item 21 was part of a three-item set that asks students to identify whether or not each method of assigning subjects to treatment was an appropriate method of random assignment. The percent of students who correctly identified that assigning students to treatments in the order that they enter a classroom is not an appropriate method of random assignment increased by about 20 percentage points from pretest to posttest. For item 22, almost 80% of students on the pretest correctly identified a research question that would be appropriate for a randomized experiment, and this increased to over 90% on the posttest.

4.3.6.3 Items with non-statistically significant increases in percentages of students with correct responses from pretest to posttest

For four items, the percent of students who answered correctly on the pretest was less than 80% and increased from pretest to posttest, but the increases were not statistically significant. The percent of correct responses for each of these items, along with their *p*-values from the McNemar's tests, are shown in Table 4.18 below. Two items (8 and 9) were from the sampling section, and the other two items (17 and 19) were from the assignment section.

Table 4.18
Items with non-significant gain from pretest to posttest, percent correct less than 80% on pretest

Item	Measured Learning Outcome	*n*	% of Students Correct		McNemar's test *p*
			Pretest	Posttest	
8	Ability to understand that sample statistics vary from sample to sample.	125	65.6	73.6	.121
9	Ability to recognize that random sampling is the most salient issue when using a sample to generalize to the population.	125	51.2	64.8	.016
17	Ability to understand how a confounding variable may explain the association between an explanatory and response variable.	125	67.2	80.0	.011
19	(One of three-item set): Ability to understand that random assignment is the best way to balance out groups with respect to confounding variables.	122	65.0	74.4	.126

For item 8, students were asked to identify the correct explanation for why two researchers using random samples of size 25 obtained different estimates for a sample mean. Over half of students correctly identified on pretest and posttest that the sample means varied because each sample represented a different subset of the population. However, on pretest and posttest, over 20% of students incorrectly selected the choice that the sample means varied because they were from small samples (see Appendix J).

Item 9 asked students to identify the main problem with printing a headline that makes a generalization statement from a study that used a convenience sample. On the pretest, about half of students correctly recognized that the main problem was the lack of

random sampling, while almost 30% indicated this was because the sample size was too small. On the pretest, the least frequent answer choice (under 10% of students) was identifying the main problem as the lack of random assignment (thus indicating possible confusion between random sampling and random assignment; see Appendix J). On the posttest, about 65% of students chose the correct answer, but the percent of students incorrectly identifying the lack of random assignment as the main problem increased to over 20%. However, the percent of students who indicated that the sample size was too small decreased from 22% to 6% from pretest to posttest.

Item 17 involved being able to identify that a confounding variable could explain an association found in an observational study. The percentage of students who correctly identified this increased by about 13 percentage points from pretest to posttest. The students who chose the incorrect options were approximately evenly split between incorrectly indicating a causal claim could be made, and incorrectly saying that valid conclusions could not be drawn due to the small sample size (see Appendix J).

Item 19 was part of a three-item set that asked students to identify whether or not different methods were an appropriate way to randomly assign subjects to treatments. For this item, over 70% of students on pretest and posttest correctly recognized that having students self-select groups, and then randomly assign treatments to groups, was not a valid method of random assignment for balancing out potential confounding variables.

4.3.7 Pretest to posttest changes in IDEA item sets

The IDEA instrument contained three different item sets: Items 1-2 about identifying the sample and population, items 12-15 about distinguishing between association-only statements and causation statements, and items 19-21 about identifying

appropriate and inappropriate ways to randomly assign subjects to treatments (see blueprint in Appendix I). Response patterns for each of these sets of items were analyzed. Alluvial plots were created in order to visualize the changes in number of correct responses to each item set from pretest to posttest.

4.3.7.1 Response patterns for items 1-2

Items 1-2 were in the sampling portion of the IDEA assessment, and involved identifying the sample (item 1) and the population to which inferences could be made (item 2) from a study. For this two-item set, Table 4.19 and Figure 4.2 display the changes in number of items answered correctly from pretest to posttest.

Table 4.19
Number of correct responses (and percent of n = 125 responses) on pretest and posttest for items #1 and #2

	Number of items correct on posttest			
Number of items correct on pretest	0	1	2	Total
0	1 (0.8%)	7 (5.6%)	4 (3.2%)	12 (9.6%)
1	1 (0.8%)	26 (20.8%)	35 (28.0%)	62 (49.6%)
2	4 (3.2%)	8 (6.4%)	39 (31.2%)	51 (40.8%)
Total	6 (4.8%)	41 (32.8%)	78 (62.4%)	125 (100%)

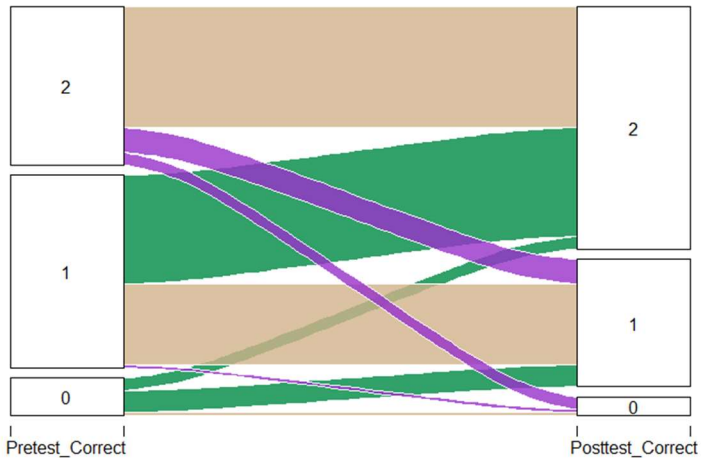


Figure 4.2. Alluvial plot for items about identifying sample and population (items 1-2)

On the pretest, almost half of students only answered one item correctly, while just over 40% answered both items correctly. On the posttest, the number of students who answered both items correctly increased to slightly over 60%. Just over one-third of students increased in the number of correct items from pretest to posttest, and about one-half of students maintained the same performance (answered the same number of items correctly on the pretest and posttest). The great majority of students were able to either correctly identify the sample (item 1), identify the population (item 2), or identify both population and sample on both pretest and posttest.

4.3.7.2 Response patterns for items 12-15

Items 12-15 were in the assignment portion of the IDEA assessment. These four items consisted of headline statements, and students were asked to identify whether the statement only made an association, or implied causation. For this four-item set, Table 4.20 and Figure 4.3 display the changes in number of items answered correctly from pretest to posttest.

Table 4.20
Number of correct responses (and percent of n = 125 responses) on pretest and posttest for items #12-15

	Number of items correct on posttest					
Number of items correct on pretest	0	1	2	3	4	Total
0	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.8%)	1 (0.8%)
1	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.8%)	1 (0.8%)
2	0 (0.0%)	1 (0.8%)	2 (1.6%)	1 (0.8%)	6 (4.8%)	10 (8.0%)
3	0 (0.0%)	1 (0.8%)	0 (0.0%)	5 (4.0%)	9 (7.2%)	15 (12.0%)
4	0 (0.0%)	0 (0.0%)	5 (4.0%)	10 (8.0%)	83 (66.4%)	98 (78.4%)
Total	0 (0.0%)	2 (1.6%)	7 (5.6%)	16 (12.8%)	100 (80.0%)	125 (100%)

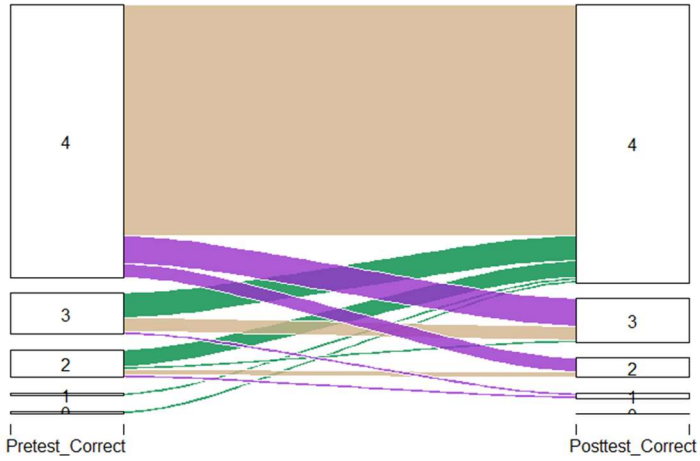


Figure 4.3. Alluvial plot for items about distinguishing association-only and causation statements (items 12-15).

About two-thirds of students answered all four items correctly on both pretest and posttest. About 70% of students maintained the same performance (answered the same number of items correctly on both pretest and posttest) for this set of items. As seen in the alluvial plot, the amount of students whose performance on the item set increased was

about the same as the amount of students whose performance decreased. Most students did well with being able to distinguish between association-only statements and causation statements, and there were very few students who answered less than half of the item set incorrectly.

4.3.7.3 Response patterns for items 19-21

Items 19-21 were in the assignment portion of the IDEA assessment. Each of these three items presented students with potential ways to assign students to four different treatment groups. Students were asked to identify whether each method was a valid way to randomly assign subjects in order to enable cause-and-effect conclusions. For this three-item set, Table 4.21 and Figure 4.4 display the changes in number of items answered correctly from pretest to posttest.

Table 4.21
Number of correct responses (and percent of n = 125 responses) on pretest and posttest for items #19-21.

	Number of items correct on posttest				
Number of items correct on pretest	0	1	2	3	Total
0	0 (0.0%)	2 (1.6%)	2 (1.6%)	3 (2.4%)	7 (5.6%)
1	1 (0.8%)	3 (2.4%)	7 (5.6%)	12 (9.6%)	23 (18.4%)
2	3 (2.4%)	5 (4.0%)	12 (9.6%)	22 (17.6%)	42 (33.6%)
3	0 (0.0%)	5 (4.0%)	6 (4.8%)	42 (33.6%)	53 (42.4%)
Total	4 (3.2%)	15 (12.0%)	27 (21.6%)	79 (63.2%)	125 (100%)

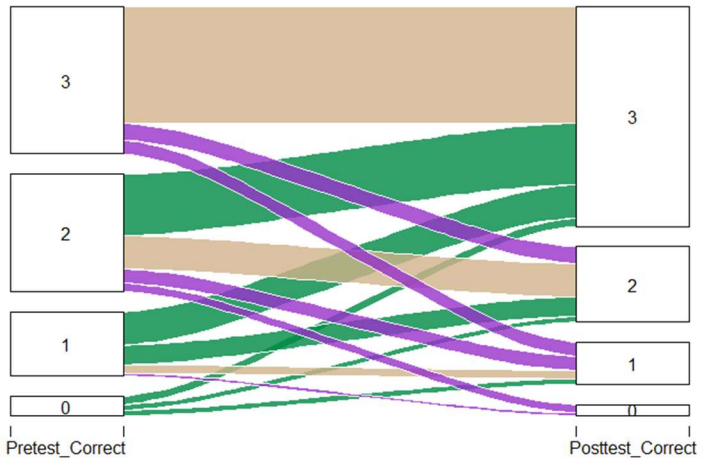


Figure 4.4. Alluvial plot for items about identifying appropriate methods of random assignment to treatments (items 19-21).

For items 19-21, the percent of students who answered all three items correctly increased from about 40% to 60% from pretest to posttest. On the pretest, about three-quarters of students answered either two or three items correctly, and this amount increased to about 85% for the posttest. For this set of items, almost half of students maintained the same performance (answered the same number of items correctly on both pretest and posttest). The alluvial plot shows that there were more students who increased their performance from pretest to posttest than students who decreased their performance. On the posttest, most students were successfully able to identify the two incorrect methods of random assignment and the one correct method.

4.3.8 Summary of quantitative analyses

The IDEA test was administered as a pretest and posttest. Overall, there were significant increases in performance when looking at total score, and also when looking at the sampling and assignment subscores, with an effect size of approximately one standard

deviation for each. Changes from pretest to posttest scores were not significantly different by section.

When examining changes in individual items more closely, there were no items with a significant decrease in percent correct from pretest to posttest. On both pretest and posttest, students appeared to do well identifying the sample, determining whether a study is observational or experimental, distinguishing between association and causation statements, and identifying that random assignment is the best design for answering research questions about causation. Some of the learning goals for which students showed learning gains were understanding what it means to make an appropriate generalization, understanding that a small, random sample is preferable to a larger, biased sample, and recognizing that correlation does not imply causation. On both pretest and posttest, students appeared to overemphasize sample size over sampling method. Many of them identified that a sample size of 500 was too small to generalize, despite the fact that it was taken randomly. Students showed modest, but not significant, gains in ability to understand that sample statistics vary from sample to sample, ability to recognize that random sampling is the most salient issue when using a sample to make a claim about a population, and ability to understand how a confounding variable may explain associations between explanatory and response variables.

4.4 Results from qualitative analysis of open-ended assessments

Two open-ended assessments were created as a part of this study, a group quiz and a lab assignment. The 128 lab assignments and 43 group quizzes were analyzed qualitatively according to the coding scheme described in section 3.9.2 and Appendix L. The percent of assignments that were labeled with each code was recorded for each section separately and

for all sections overall. This section presents the results found from these assessments beginning with inter-rater agreement between the researcher and graders using the scoring rubrics. Then, the subsections that follow present coding tables for each of the categories of codes. The full set of coding tables, as well as the coding categories and their abbreviations, can be found in Appendix M.

4.4.1 Inter-rater agreement

In order to examine the fidelity of rubric implementation, the researcher obtained ungraded copies of the quizzes and labs, and then graded them independently. The group quiz was graded by the teaching assistant of each of the in-class sections. For the online section, the instructor graded the group quizzes. The graders used the rubric given by the researcher (Appendix F2), and were asked to contact the researcher with any questions while they graded. The graders did not contact the researcher with any questions during the grading process.

To stay consistent with how labs were typically scored in the course, the researcher created a lab rubric for holistic scoring on a scale of 0 to 3, emphasizing the main points that students should understand (see Appendix G2). The instructors of the three in-class sections graded their students' labs. For the online section, the teaching assistant graded the lab assignments. Again, the graders were asked to contact the researcher with any questions while grading, but no questions were asked during the process.

4.4.1.1 Inter-rater agreement: Group quiz

The group quiz was scored on a scale of 0 to 6 points according to the quiz rubric (see Appendix F2). Ungraded copies of the 43 group quizzes were scored independently

by the researcher, and then the researcher's scores were compared to those given by the graders.

The rubric for the grading of the quiz included only whole-point or half-point possible deductions, but some graders took off quarter points for certain mistakes. Seven total group quizzes included quarter-point deductions. Since the rubric did not describe quarter-point deductions, these quizzes were examined further along with the rubric, and the graders' scores that had included quarter-points were rounded to the nearest half point depending on the rubric. For example, in question 6 (see quiz 5 rubric in Appendix F2), two groups indicated that a headline making an association was all right because the study was observational, but failed to mention that the headline was making a generalization, allowable by the random sampling. The graders took off a quarter point for the two groups, but based on the rubric, the grade was revised to take half a point for failing to address the generalization. Another group answered question 6 correctly by saying that the headline making a generalization was appropriate due to random sampling, but wrote the incorrect phrase "the word 'associated' implies a generalization about a population." Since this statement is incorrect, the grader took off a quarter point. The rubric had not specified taking off points for minor phrasing errors when the rest of the answer was correct, so this group's grade was revised to round up to the nearest half point.

After these revisions, the polychoric correlation between the researcher's score and the grader's score for the 43 quizzes was computed to be 0.98, indicating a high level of agreement. Only two group quizzes showed discrepancies between the grader and the researcher. In one quiz, the students misinterpreted a claim in question 1 as being a causal claim, but otherwise mentioned that random assignment was necessary for the causal claim.

The rubric addressed this possibility and instructed the grader to take off a half point for this mistake, but the grader (a teaching assistant) took off a full point. This resulted in the group's quiz being scored a 5 by the grader and a 5.5 by the researcher. In the other quiz, in question 6, students wrote that an association claim was appropriate because the study did not include random assignment, but failed to recognize that the claim was making a generalization to a population. The grader did not take points off for this, but the researcher took off a half point as the rubric instructed for this mistake. This resulted in the group's quiz being scored a 5.5 by the grader and a 5 by the researcher.

4.4.1.2 Inter-rater agreement for the lab assignment

The lab assignment was scored holistically, with each lab receiving a total score from 0 to 3 points according to the lab rubric given to the graders by the researcher (see Appendix G2). For sections 1, 2, and 3, the instructor graded the lab assignments. For section 4, the teaching assistant graded the lab assignments. There were 132 total labs submitted by students and graded independently by the researcher, and then the researcher's scores were compared to those given by the graders. Table 4.22 shows the distribution of the researcher's scores and the grader's scores. As there is some degree of subjectivity to assigning holistic scores, there was disagreement in scoring for 39 out of the total 132 lab assignments (about 30% of assignments). However, most of the disagreements were of a half point or less, and the discrepancy was never larger than 1 point. As these holistic scores are ordinal, a polychoric correlation between the grader's score and the researcher's score was computed to be 0.80.

Table 4.22
Lab scores given by grader and researcher

	Researcher's score							
		0.5	1	1.5	2	2.5	2.75	3
Grader's score	0.5	**3**	0	0	0	0	0	0
	1	1	**5**	1	0	0	0	0
	1.5	2	4	**7**	1	0	0	0
	2	0	1	2	**5**	4	0	1
	2.5	0	0	1	2	**16**	2	2
	2.75	0	0	0	0	4	**3**	1
	3	0	0	0	1	3	6	**54**

Note. Diagonal cells are bolded to show the assignments for which there was agreement in scores between the grader and researcher.

4.4.2 Results from qualitative analysis of lab assignment

The lab was an individual assignment completed at the end of the unit, and involved reasoning about conclusions that can be made from each of two different studies involving infants and peanut allergies. The lab assignment can be found in Appendix G1. Table 4.23 shows the percent of lab assignments that were labeled as falling into each coding category of incorrect thinking, correct thinking, and ambiguity (see coding scheme described in section 3.9.2 and Appendix L).

Table 4.23
Percent of students displaying behaviors in each coding category for lab assignment

		% of Section				
Code	Behavior	1 (n = 40)	2 (n = 31)	3 (n = 27)	4 (n = 30)	% of all (n = 128)
[I]	*Misconceptions/Incorrect Thinking*					
[I-TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	15.0	22.5	33.3	30.0	24.2
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	2.5	0.0	0.0	13.3	3.9

Code	Behavior	% of Section 1 (n = 40)	2 (n = 31)	3 (n = 27)	4 (n = 30)	% of all (n = 128)
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	15.0	12.9	14.8	20.0	15.6
[C]	*Correct Thinking*					
[C-SG]	*Makes connections between sampling and generalization: Either mentions lack of RS OR how sample is different from population*[a] *(at least one C-SG code)*	100.0	100.0	96.3	70.0	92.2
[C-AC]	*Makes connections between random assignment and causation: Either mentions lack of RA OR how groups are different from each other (confounding)*[b] *(at least one AC code)*	95.0	93.6	96.3	66.7	88.3
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	57.5	51.6	48.2	23.3	46.1
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	22.5	38.7	22.2	20.0	25.8
[A]	*Ambiguity (at least one A code)*	5.0	9.7	14.8	20.7	11.8

Note. RS refers to random sampling, and RA refers to random assignment.

Over 90% of students successfully connected random sampling to generalization. Both of the studies described in the lab assignment were conducted using convenience samples. Almost all students successfully explained that the results could not be generalized to a defined population of infants, either by pointing out the lack of random sampling, or pointing out differences in characteristics between the infants in the convenience sample and a broader population of infants. The online section, however, had a lower percentage of labs (70%) demonstrating correct understanding of sampling and generalizability, unlike the other sections which had more than 90% of students demonstrating this correct understanding.

Nearly 90% of students overall successfully connected random assignment to causal claims. One of the studies described in the lab used random assignment, while the other did not. Successfully connecting random assignment to making causal claims involved either identifying that random assignment helped to enable causal conclusions, or identifying that confounding variables could make groups different from each other and thus impede causal claims. Again, the online section performed worse on this than the other sections. Only about two-thirds of online students successfully made this connection between random assignment and causal claims compared to more than 90% for all of the other sections.

On the lab assignment, students were not asked to elaborate specifically about why random sampling is connected to generalization and why random assignment is connected to causation. However, approximately half of students in each in-class section and one-fifth of students in the online section included more detail about either why random sampling helps with generalizability, or why random assignment helps with causal claims. In general, more students elaborated about random assignment and causal claims than about random sampling and generalization (see Appendix M1).

About 25% of students overall added extraneous information to their correct answers, for example addressing the lack of ability to generalize in addition to addressing causal claims, when the question was only asking about causal claims. In general, students were more likely to bring in extraneous information about generalizability when the question was about causal claims, than they were to bring causal claims extraneously into a question about generalizability (see Appendix M1).

Just over 10% of students had answers that were ambiguous as to whether they were reasoning correctly about the appropriate randomization in the study design, and the online section tended to do this more frequently than the in-class sections. These students could have been reasoning correctly, but their answers were sometimes not specific enough to indicate whether they were connecting the correct study design with the correct conclusion (for example, referring to "random" study designs without specifying random sampling or random assignment).

Almost 25% of students demonstrated misunderstandings on their lab assignment regarding which study designs help with generalization or causation. As seen in Appendix M1, the most common error shown was bringing up random assignment, and not random sampling, when the question was about generalizing to a population. Online students also tended to have more problems than in-class students with incorrectly discussing only random sampling when the question was about making causal claims.

Few students overall gave answers showing misconceptions related to sample size, but 10% of online students did show incorrect thinking about the role of sample size, such as suggesting that one can generalize only due to the large sample size (Appendix M1).

The study descriptions given in the lab assignment were taken from real journal articles, and about 15% of students overall showed difficulty understanding from the study description whether random sampling and/or random assignment were used. For example, some students assumed random sampling was done, even though the study descriptions specified that the subjects were "recruited" and did not mention random sampling.

Performance on lab question about recognizing confusion between random sampling and random assignment

The last two lab questions were examined separately with specific codes. The penultimate question (question 13) asked students whether a classmate was correct in stating that if random sampling had been done in the study, one could make causal claims about peanut consumption and allergies. Table 4.24 shows the distribution of students' labs falling into each of the coded behaviors.

Table 4.24
Percent of students displaying behaviors for each code for question 13

		% of Section				
Code	**Behavior**	1 (n = 40)	2 (n = 31)	3 (n = 27)	4 (n = 30)	% of All *(n = 128)*
I-LAB13-RSCC	*Says classmate is correct that RS leads to causation*	5.0	6.5	0.0	23.3	8.6
[C-LAB13]	*at least one C "correct" code for question #13: Either explains RS is only for generalization (C-LAB13-RSGEN), or explains need for RA for causation[c] (C-LAB13-RACC)*	87.5	87.1	96.3	56.7	82.0
C-LAB13-RSGEN	*Says RS is only for generalization*	52.5	71.0	55.6	26.7	51.6
C-LAB13-RACC	*Correctly brings up need for RA for causation (or problems with confounding)*	82.5	64.5	66.7	53.3	68.0

Over 80% of students appropriately explained that the classmate's statement was incorrect, using one of two possible approaches. The first possible approach, used by about half of students, was stating that random sampling was not for causation, but for generalization. The second possible approach, used by almost 70% of total students, was to mention that lack of random assignment and/or confounding variables prevent causal

claims from being made. Although nearly all in-class students reasoned correctly about this question, only a little over half of online students reasoned correctly, with more than 20% of them incorrectly saying that the classmate's statement was right.

Performance on lab question about using study results to make decisions

The final question on the lab assignment involved a hypothetical colleague wondering if she should avoid peanuts during pregnancy, based on findings of a study that found a link between frequent peanut consumption during pregnancy and a higher incidence of peanut allergies. The study was observational, using recruited subjects from a convenience sample. A correct answer involved recognizing the design limitations of this study which could impede generalization and/or causation. Students could also have correctly argued that it is unrealistic to make decisions based on only one study, but no students did this. Students were asked to reason about this question "based on the design of this study," and they could do this either by addressing the lack of ability to generalize to a population that would definitely include the colleague, or by addressing the lack of ability to make causal claims between peanut consumption during pregnancy and an incidence of allergies, or both. Table 4.25 shows the distribution of students' labs falling into each of the coded behaviors for the final question on the assignment.

Table 4.25
Percent of students displaying behaviors for each code for question 14.

		Section				All
		1	2	3	4	
		(*n* = 40)	(*n* = 31)	(*n* = 27)	(*n* = 30)	(*n* = 128)
Code	**Behavior**					
[C-LAB14]	*Either mentions lack of ability to make causal claims, or lack of ability to make generalizations (at least one C code)*	90.0	83.8	85.2	50.0	78.1
C-LAB14-NOCC	*Mention lack of ability to make causal claims*	80.0	83.9	85.2	36.7	72.9
C-LAB14-NOGEN	*Mention lack of ability to generalize*	52.5	32.3	51.9	33.3	42.3
I-LAB14-PVAL	*Decision based only on p-value*	2.5	6.5	3.7	6.7	4.7
I-LAB14-NOSD	*Decision based on factors not related to study design or results*	7.5	6.5	7.4	20.0	10.2

Over 75% of students correctly discussed either the lack of ability to generalize, or the lack of ability to make causal claims as a limitation of the study. Some students did both; for example, everyone in sections 2 and 3 who wrote about lack of generalization also wrote about the lack of ability to make causal claims. Only half of online students correctly reasoned about study design limitations, compared to more than 80% in each of the three in-class sections. In general, students were more likely to critique the lack of ability to make causal claims than the lack of ability to generalize.

A handful of students gave incorrect answers, such as making a decision based only on the low *p*-value. For example, a few students said they would advise their colleague to avoid peanuts because a significant association between peanut consumption during pregnancy and higher incidence of allergies had been found. Other students, especially

many in the online section, gave answers that did not refer to the study design at all. For example, students sometimes used their own contextual knowledge about the study, such as claiming that the decision depends on whether the colleague decides to breastfeed or not.

4.4.3 Results from qualitative analysis of quiz questions involving news headlines

The quiz consisted of three different contexts, two of which involved interpretation of whether certain news headlines were appropriate given the way in which studies were designed. These two contexts will be discussed in this subsection, and the other context (involving an experiment) will be discussed in subsection 4.4.4. Questions 1 and 2 involved a Gallup-Healthways survey about alcohol consumption and emotional health (henceforth referred to as the "Gallup" questions). Questions 5 and 6 involved a hypothetical study about GPA predicting admission to medical schools in the United States (henceforth referred to as the "admissions" questions). In both of these scenarios, random sampling was used, but random assignment was not.

The same codes were used for these two sets of questions as were used to code the lab, except for the codes that were specific to the last two lab questions, and the code *C-SG-CHAR* which involved pointing out that the sample was likely not representative of the population (incorrect here as random sampling was done). Also, two codes were added to represent difficulties interpreting when headlines were making a generalization and/or causal claim. Table 4.26 and Table 4.27 show the distribution of group quizzes falling into each of the coded behaviors for the Gallup questions and for the admissions questions, respectively.

Table 4.26

Percent of students displaying behaviors in each coding category for Gallup questions

Code	Behavior	% of groups per section				% of all groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
[I]	*Misconceptions/Incorrect Thinking*					
[TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	14.3	8.3	0.0	25.0	11.6
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	7.1	8.3	0.0	0.0	4.7
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	7.1	0.0	0.0	0.0	2.3
[C]	*Correct Thinking*					
C-SG-RSGEN	Recognizes that random sampling is relevant for generalization (in this case, we have a random sample so we can generalize to a population)	78.6	83.3	100.0	75.0	83.7
[C-AC]	*Makes connections between assignment and causation. Either mentions lack of RA OR how groups are different from each other (confounding) (at least one AC code)*	92.9	83.3	77.8	100.0	88.4
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	21.4	0.0	0.0	37.5	14.0
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	7.1	33.3	44.4	87.5	37.2
[A]	*Ambiguity (at least one A code)*	0.0	8.3	22.2	0.0	7.0
Quiz-specific codes for items involving headlines						
I-QUIZ-HGEN	Not recognizing when headline is/is not making a generalization	21.4	8.3	0.0	12.5	11.6
I-QUIZ-HCC	Not recognizing when headline is/is not making a causal claim	14.3	16.7	22.2	0.0	14.0

Note. RS refers to random sampling, and RA refers to random assignment.

Table 4.27

Percent of students displaying behaviors in each coding category for admissions questions

Code	Behavior	% of groups per section				% of all groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
[I]	*Misconceptions/Incorrect Thinking*					
[TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	0.0	0.0	11.1	12.5	4.7
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	0.0	0.0	0.0	12.5	2.3
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	0.0	0.0	0.0	0.0	0.0
[C]	*Correct Thinking*					
C-SG-RSGEN	Recognizes that random sampling is relevant for generalization (in this case, we have a random sample so we can generalize to a population)	78.6	66.7	77.8	50.0	69.8
[C-AC]	*Makes connections between assignment and causation. Either mentions lack of RA OR how groups are different from each other (confounding) (at least one AC code)*	92.9	100.0	77.8	87.5	90.7
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	0.0	0.0	11.1	0.0	2.3
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	42.9	91.7	33.3	25.0	51.2
[A]	*Ambiguity (at least one A code)*	7.1	0.0	22.2	0.0	7.0
Quiz-specific codes for items involving headlines						
I-QUIZ-HGEN	Not recognizing when headline is/is not making a generalization	21.4	41.7	0.0	50.0	27.9
I-QUIZ-HCC	Not recognizing when headline is/is not making a causal claim	0.0	0.0	0.0	0.0	0.0

Note. RS refers to random sampling, and RA refers to random assignment.

For both the Gallup and admissions scenarios, the majority of student groups correctly reasoned about the appropriateness of a generalization headline (given that the data came from a random sample) and the inappropriateness of a causation headline (given that both studies were observational). For both scenarios, about 90% of all student groups correctly explained that causal headlines were not appropriate, either because random assignment was not present, or because other variables could explain the associations found. However, for the Gallup context, over 80% of groups correctly reasoned that the random sampling made the generalization headline appropriate, compared to just under 70% for the admissions questions. Online students performed somewhat worse than the other sections on the generalizability question of the admissions scenario, with only half of them correctly identifying that the generalization headline was not supported by the study design.

One behavior that was common in the admissions scenario, especially among section 2, was bringing in extraneous information, talking about generalization when the question was about causation, and vice-versa. About half of all groups did this, including over 90% of groups in section 2. However, groups who brought extraneous information still spoke correctly about the study design relevant to the question. Student groups tended to address the lack of ability to make causal claims when being asked about the headline that made a generalization. For example, various groups stated that the word "association" in the headline in question 2 (Appendix G1) was fine because it did not make a causal claim, and also that the generalization "American adults" in the headline was appropriate due to the random sampling. In the Gallup questions, only about 37% of groups brought in extraneous information, but this behavior was very common in the online section (87.5%

of groups). These groups tended to address both generalization and causation when asked about each headline, but wrote about both correctly.

In both sets of items, none of the questions explicitly asked students to elaborate on why each study design allowed for each conclusion. For example, merely stating that the random sampling made it possible to create a headline generalizing to the population was acceptable, and students were not asked to explain why (e.g., describing that random sampling helps to avoid bias by making every subject equally likely to be selected). Still, about 14% of groups answering the Gallup questions explained why the random sampling allowed for generalization or why the lack of random assignment (or presence of confounding variables) did not allow for headlines making causal claims. However, only student groups from sections 1 and 4 provided these types of more complete answers for the Gallup scenario. For the admissions scenario, only one group (from section 3) out of all student groups explained why the observational study did not allow for a causal claim between GPA and medical school admissions.

Misunderstandings relevant to sample size and difficulties understanding study descriptions were rare for both contexts. (The sample size for the Gallup survey was over 300,000, and for the admissions study it was 250, and both used the words “random sample” to describe the sampling method.) Incorrect answers about which study designs help with which types of conclusions were slightly more common in the Gallup context (about 12%) than in the admissions context (less than 5%). Also, answers that made it ambiguous whether students were correctly reasoning about the questions were not common, with about 7% of groups demonstrating this behavior for each scenario.

The most common problem among incorrect answers, especially for the admissions context, was failing to recognize when a headline was making a generalization. For each of the Gallup and admissions item sets, about 20% of student groups gave answers that showed a failure to recognize that a headline made a generalization. Class observation notes stated that when examining the headline "In U.S., Moderate Drinkers Have Edge in Emotional Health," students tended to focus more on the term "edge" while overlooking the phrase "in U.S." Thus, they wrote that this only implied an association, so the headline was appropriate, without addressing the fact that the headline was also making a generalization. In the Gallup context questions, about 14% of groups had trouble recognizing when a causal claim was being made, but this problem was not present in the admissions context.

4.4.4 Results from qualitative analysis of quiz questions involving experimental study

The other scenario on the quiz (questions 3 and 4) involved a context in which an experiment was conducted to examine the effect of bowl size on amount of ice cream served. The subjects were nutritionists in Massachusetts at an ice cream social. Students were first asked if it was likely that factors other than bowl size could explain differences in amount served (question 3), and then whether the results were generalizable to all nutritionists in Massachusetts (question 4). Questions 3 and 4 will henceforth be referred to as the "ice cream" questions.

The same codes were used for these two questions as were used in coding the lab, except for codes specific to the last two lab questions, and the code *C-AC-CONFV* which involved pointing out that confounding variables likely make groups differ from each other and would not constitute correct reasoning in this context where random assignment was

used. The ice cream questions did not involve examining headlines, so two codes directly related to interpreting headlines (*I-QUIZ-HGEN* and *I-QUIZ-HCC*) that were used in the other two quiz scenarios were not used here. Table 4.28 shows the distribution of group quizzes falling into each of the coded behaviors for the ice cream questions.

Table 4.28
Percent of students displaying behaviors in each coding category for ice cream questions

		% of groups per section				% of all
Code	Behavior	1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	Groups (n = 43)
[I]	*Misconceptions/Incorrect Thinking*					
[TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	35.7	25.0	0.0	37.5	25.6
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	7.1	0.0	0.0	12.5	4.7
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	7.1	33.3	33.3	37.5	25.6
[C]	*Correct Thinking*					
[C-SG]	*Makes connections between sampling and generalization: Either mentions lack of RS OR how sample is different from population (at least one SG code)*	92.9	83.3	88.9	87.5	88.4
C-AC-RACC	Recognizes that random assignment is relevant for causation (in this case, we have random assignment so we can make causal claims)	85.7	58.3	77.8	87.5	76.7
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	71.4	33.3	77.8	62.5	60.5
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	14.3	16.7	11.1	12.5	14.0
[A]	*Ambiguity (at least one A code)*	0.0	8.3	0.0	12.5	4.7

Note. RS refers to random sampling, and RA refers to random assignment.

For the ice cream questions, almost 90% of groups overall recognized the lack of generalizability to all nutritionists in Massachusetts, either by citing the lack of random sampling, or stating that the nutritionists were at an ice cream social and may not represent all nutritionists in the state. All sections performed reasonably well on the generalizability question. About three-quarters of all groups correctly answered that it is not likely that factors other than bowl size may explain any differences in average amount of ice cream served, recognizing that the lack of assignment should theoretically balance out these other factors. However, section 2 appeared to perform worse on this question than the other three sections, with only 58% of student groups reasoning correctly about confounding variables theoretically being balanced out by the random assignment.

Unlike the other two quiz scenarios, for the ice cream scenario, more than half of groups went into depth about either why random assignment is linked to causal claims, or why random sampling is linked to generalization, even though providing an explanation was not directly prompted by the questions. In question 3, many groups explained why random assignment made it unlikely that other factors could explain differences in bowl size, explaining how it should balance out confounding variables between the groups. However, section 2 appeared considerably less likely to explain this (with only 33% of groups) than the other three sections. Fewer than 20% of groups explained why the lack of random sampling did not allow for generalization (see Appendix M2, questions 3 and 4). About 14% of groups brought in extraneous information about random assignment or causation, while still talking correctly about generalizability in question 4. For example, many answers to question 4 stated that while this convenience sample does not allow generalization to all Massachusetts nutritionists, the random assignment does allow for

causal claims to be made. Only a few groups (less than 5%) wrote answers that made it ambiguous whether or not they were reasoning correctly (for example, mentioning "randomness" or saying that both random assignment and random sampling are needed to make causal claims that can be generalized).

Just over 25% of student groups displayed misunderstanding about which study designs help with which types of conclusions for the ice cream questions. The most common error was stating in question 3 that the lack of random sampling (or the fact that this was a convenience sample) makes it likely that factors other than bowl size may explain differences in average amount of ice cream served. No groups in section 3 made this error, but multiple groups in each of the other sections did. Also, just over 25% of groups displayed difficulty understanding the study description, mostly failing to recognize that random assignment had been used. For example, many answers to question 3 described different confounding variables that could affect ice cream serving size, failing to acknowledge that random assignment had been used in the study.

4.4.5 Summary of results from qualitative analysis

Overall, on the lab and on each set of quiz questions, the great majority of students successfully made the appropriate connections between random sampling and generalization, and random assignment and causal claims. In general, many students made these connections without elaborating further about why each study design allows for each type of conclusion (e.g., discussing bias or confounding variables), although most questions did not prompt them to do so. Most students also successfully identified and corrected the misunderstanding that random sampling lead to causation (lab question 13),

and reasoned appropriately about the limitations of a study design when making decisions (question 14).

A considerable portion of students still demonstrated misunderstandings about study design and types of conclusions, such as mixing up random sampling with random assignment. Also, students sometimes tended to have problems recognizing generalization statements, or discerning whether or not random sampling or random assignment were used in the study design. Occasionally, students gave answers that made it ambiguous whether or not they were reasoning about study design correctly, but this was not very common. Usually, there did not appear to be noteworthy differences between sections, but occasionally, the online students tended to display incorrect thinking at a higher rate than the in-class students.

4.5 Summary of results

This chapter described the results from the classroom observations and analysis of assessments from the study design unit. Overall, students appeared to improve in their understanding of study design and conclusions. Some difficulties, such as misunderstandings about the role of sample size in a study, and difficulty distinguishing between random sampling and random assignment, still prevailed in a small, but noticeable, portion of students at the end of the unit. The following chapter offers a discussion of these results, limitations of the study, and implications for future research.

Chapter 5

Discussion

5.1 Summary of the study

This study was conducted to examine students' learning of study design and conclusions in a unit developed for an introductory statistics course. A two-and-a-half-week study design unit was created and administered in an undergraduate introductory statistics course. The unit consisted of four activities, a group quiz, and a homework assignment. In addition, the *Inferences from Design Assessment* (IDEA) forced-choice assessment was created and administered as a pretest and posttest to examine changes in students' understanding. Activities for the three in-class sections were videotaped and observed by the researcher and a co-observer, and the researcher examined the online class discussion forums.

Based on a review of literature and introductory statistics textbooks, a test blueprint was developed for the IDEA pretest and posttest, and activities and open-ended assessments were created. The activities were either created or modified from previous course activities (Zieffler et al., 2013). Activities were first modified based on several rounds of feedback from the two advisors on this project, and then were further modified after feedback from all three instructors of the course in which they would be implemented. Open-ended assignments and rubrics for these assignments were created and modified based on feedback from advisors and instructors. Lesson plans for the instructors and observation forms for the observers were also created.

The IDEA test was created by taking or modifying items from existing assessments in statistics education (e.g., CAOS, delMas et al., 2007; ARTIST, Garfield et al., 2002) that

would fit each of the learning goals on the blueprint. After initial feedback from the advisors on this project, the blueprint and IDEA test were sent to three external reviewers, all experts in statistics education. IDEA was further modified based on their feedback, then placed online for students to take prior to the study design unit, and then again after the unit.

Class observation notes, students' IDEA pretest and posttest answers, and all responses on the group quiz and homework assignment were examined in order to explore students' conceptual understanding of random sampling, random assignment, and the role that these designs play in conclusions that can be made from studies. Quantitative analyses were conducted on the IDEA test to examine changes from pretest to posttest. Qualitative analyses using a coding scheme were conducted on the group quiz and homework assignment.

5.2 Synthesis of the results

Prior to this study, there has been limited research on introductory statistics students' understanding of the purposes of random sampling and random assignment. This study was developed with the goal of answering the research question: *How does introductory statistics students' conceptual understanding of study design and conclusions (in particular, unbiased estimation and establishing causation) change after participating in a learning intervention designed to promote conceptual change in these areas?* This section offers a discussion of the study's contributions to research about students' learning of random sampling, random assignment, and conclusions that can be made from each.

5.2.1 Students' prior knowledge

It is important to take into account students' prior knowledge, as it plays an important role in how students experience new concepts (Vosniadou & Brewer, 1987; Vosniadou, 2013). Prior to the study design unit, students had already studied the following topics in the course:

- Randomness (including human intuitions about randomness and modeling random behavior)
- Strength of evidence as measured by *p*-values
- Bootstrap interval estimates
- Randomization tests
- Features of distributions
- Effect sizes

Data from the IDEA pretest suggests that students came into the unit already having considerable understanding of some of the learning goals. There were 9 IDEA items (representing 6 different learning goals out of the 16 total learning goals) with very high performance (more than 80% correct) on both pretest and posttest. Given students' prior experience in the course working with samples, it is not surprising that one of the learning goals represented by these items was identifying a sample (item 1). Although they had not learned about random sampling yet, most students also appeared to understand on the pretest that random sampling is preferable to non-random methods of sampling (item 7). Given all of the prior course activities and homework assignments involving data from randomized experiments, it is not surprising that most students on the pretest could distinguish between an observational and an experimental study (item 10), indicate that

randomly assigning individuals to groups is the best way to balance out confounding variables (item 20), recognize that a randomized experiment is needed to answer questions about causation (item 11), and distinguish between association and causation statements (items 12-15).

Students' prior knowledge outside of taking a statistics course may also play a role in their high performance on these items. For example, students who have a good grasp of the English language and have experience with reading headlines that make causal claims may find it easy to recognize terms that make a causal statement such as "leads to" or "improves," which would help them with items 12-15. Also, students' contextual knowledge can guide their answers (Wroughton et al., 2013) and this prior knowledge could have been informative on the pretest. For example, students could have encountered information about drug trials in the media and may already have exposure to the idea that a randomized experiment is needed to conclude whether a vitamin supplement is effective, thus guiding their answer to item 11.

It is important to recognize that students came into the study design unit with a considerable amount of prior knowledge already, especially regarding experimental study design, after having worked with several contexts involving randomized experiments. The results from the analysis of IDEA might have been different if they had not already come in with this prior knowledge. In the following sections, performance on some IDEA items is compared with some similar items from previous assessments (e.g., CAOS, delMas et al., 2007). However, it is important to consider that the items went through changes before becoming a part of IDEA, and although IDEA was taken immediately after the study design

unit, other assessment data used for comparison involves students taking a posttest at the end of the course.

5.2.2 Areas of success

Based on results from the IDEA test and open-ended assessments, there are various learning areas in which students appeared to improve or do well by the end of the curriculum. This subsection will discuss those learning areas.

5.2.2.1 Sampling and generalization

There were five learning goals related to sampling and generalization on the IDEA test for which students showed statistically significant improvement. Students significantly improved in their ability to identify a population to which inferences can be made (item 2), even though they did not do as well with this learning goal as they did in identifying the sample (item 1). This is not surprising, as they had worked with samples many times earlier in the course, but had not spent much time discussing the populations that those samples represent. Students also significantly improved in understanding what it means to make an appropriate generalization (item 3), and identifying factors that allow a sample of data to be representative of the population (item 4). Students significantly improved in their ability to identify when sample estimates may be biased (item 5), although their performance on the posttest for this item is comparable to the performance on a similar item for students who took a pervious iteration of the CATALST curriculum (Sabbag, 2013). Students also significantly increased in their ability to identify that a small, random sample is preferable to a large, biased sample (item 6). This finding is encouraging, as the curriculum was designed to help target the incorrect idea that larger samples are always preferable to smaller ones.

There were two learning goals on the IDEA test related to sampling and generalization for which students showed modest (not statistically significant) improvement. One of these learning goals involved understanding that statistics vary from sample to sample (item 8). This is an idea that students had already seen in the curriculum when studying bootstrapping, but perhaps was reinforced in the study design activities involving sampling. On item 8, the most popular distractor option for both pretest and posttest was that the sample means varied because they were computed from small samples. One of the expert reviewers of the assessment suggested that this distractor had some truth to it, but because sample means from large samples also vary, the item was kept as is with the correct answer being "the sample means varied because each sample is a different subset of the population." Still, about one-fifth of students chose this distractor option on the posttest, so perhaps this indicates students are showing correct reasoning that sample means from small samples vary more than sample means from large samples.

The other learning goal that showed non-significant improvement involved the ability to recognize that random sampling is the most salient issue when attempting to generalize to a population (item 9). This item involved recognizing a generalization statement, and then recognizing that the statement may not be accurate due to lack of random sampling. On the group quiz, students had gotten practice recognizing generalization statements and evaluating whether or not these claims could be made based on the study design. This practice may have helped students on the posttest.

On the group quiz and lab assignments, most students successfully identified whether or not results could be generalized based on reasoning about how the sample was selected. On their lab assignment, nearly all (over 90%) of students correctly mentioned

the lack of random sampling, or factors that made the sample different from the population, when asked about generalization. On the group quiz, for the Gallup poll (questions 1-2) and ice cream (questions 3-4) scenarios, over 80% of student groups provided answers with correct reasoning about sampling and generalization, whereas this percentage was just under 70% for the medical admissions (questions 5-6) scenario. Therefore, most students were able to connect issues of sampling to generalization, although they did this a bit more easily with some contexts than others.

Of interest in the coding of open-ended assessments was whether students would elaborate on why random sampling was relevant to generalization (for example, talking about avoiding bias, or about representativeness of the sample), even if the questions did not specifically ask them to elaborate. This type of elaboration might be evidence of deeper conceptual knowledge, as it would involve interrelations between pieces of knowledge about study design and conclusions (Hiebert & Lefevre, 1986; Rittle-Johnson & Alibali, 1999; Tennyson & Cocchiarella, 1986). In the original design of the quiz, some questions were phrased to elicit students to explain connections between concepts, such as asking: "Why does the random sampling in this study allow for the headline to be published?" However, the instructors of the course believed that this type of question would provide too much scaffolding, so the question was instead rephrased to ask more generally whether or not the study design supported the use of a given headline.

On the lab assignment, fewer than 20% of students overall elaborated about why the lack of random sampling in the studies described did not allow for generalization of results. On the quiz, the percentage of students elaborating about connections between sampling and generalization was also low for each of the scenarios (Appendix M2), and no

student groups elaborated on this for the medical school admissions scenario. Although students appeared to understand that random sampling is the relevant study design desired for making generalizations, they were less likely to go into detail about why this is the case. However, during classroom observations, it was noted that when instructors asked students questions during activity time to prompt them to explain why they had stated that random sampling allowed them to generalize (or why lack of random sampling did not allow for generalizations), students were generally able to explain that random sampling helped to obtain a representative sample. This suggests that perhaps students understood the deeper connections between random sampling and why it helps to make generalizations, but did not explain these connections unless specifically prompted to do so.

5.2.2.2 Assignment and causation

There were four learning goals related to assignment to groups and causation on the IDEA test which had statistically significant improvement from pretest to posttest. Two items measuring two of these learning goals have similar items on previous assessments for comparison. The first of these learning goals was the ability to understand that correlation does not imply causation (item 16), which had the largest improvement from pretest to posttest (from 28% to 78% correct). Performance on this item on the posttest was considerably better than performance on a similar item on CAOS by a national sample of introductory statistics students (delMas et al., 2007), where only just over half answered correctly on the posttest. Similarly, performance on this item on the IDEA posttest was better than performance on an administration of the CAOS test to students in a randomization-based curriculum (Tintle et al., 2012), where only around 60% of students answered a similar item correctly on a posttest and on a retention test.

The second learning goal with significant improvement was the ability to understand the purpose of random assignment in an experiment (to make groups comparable with respect to all other confounding variables), represented by item 18. This item had the second largest gain on the assessment, and a similar item on the CAOS test had previously been the most difficult item on that assessment for a national sample of introductory statistics students (about 12% on the posttest; delMas et al., 2007) and also one of the most difficult items among a sample of students in a simulation-based curriculum (under 10% correct on the posttest; Tintle et al., 2012). Over three quarters of the students who underwent the study design curriculum were able to answer this item correctly after the curriculum was implemented. Although the IDEA posttest was taken immediately after the study design unit, whereas the CAOS in previous studies was taken at the end of the course, findings suggest that students improved substantially in recognizing that correlation does not imply causation, and in understanding the purpose of random assignment, after going through this study design unit.

The third assignment item with significant improvement was item 21 from a three-item set recognizing that random assignment is the best way to balance out groups with respect to confounding variables. The fourth assignment item (22) with significant improvement involved recognizing an appropriate research question that can be answered using experimental design. These two items did not have similar items on previous assessments for comparison.

There were two learning goals related to assignment to groups and causation on the IDEA test which showed improvement that was not statistically significant. Students increased in the ability to understand how a confounding variable may explain associations

between explanatory and response variables (item 17), but the increase was not statistically significant after adjusting for multiple comparisons. Also, students improved, though not statistically significantly, in the ability to understand that randomly placing groups of people (rather than individuals) into treatment groups was not a method of random assignment that would control for the effects of confounding variables at the individual level. The study design unit did not explicitly address random assignment of groups of people, but most students were able to recognize that random assignment in an experiment should be done with individuals, if it is desired to balance out confounding variables between the individuals in each group.

Students tended to do well with reasoning correctly about random assignment and causation on the lab assignment and group quiz. On the lab assignment, almost all students made correct statements connecting random assignment to causal claims or alternatively mentioning that confounding or lack of random assignment did not allow for causal claims. On the quiz scenarios involving lack of random assignment (questions 1-2 and 5-6), nearly all student groups correctly identified that causal claims could not be made due to confounding or lack of random assignment. For the question involving the ice cream experiment (question 3), only about three-quarters of student groups correctly identified that the random assignment tended to balance out confounding variables. Note, however, that question 3 did not simply ask students whether or not causal claims could be made. Instead, it asked whether confounding variables were likely to affect the amount of ice cream served. While most student groups were able to come up with a correct answer involving the random assignment, the quiz observation notes indicated that students spent a considerable amount of time discussing this question. On some quiz papers, students

began to write down some possible confounding variables, only to then cross them out and change their answer to indicate that the random assignment makes it unlikely for confounding variables to affect the results. This suggests that perhaps, learners naturally think of confounding variables that can explain associations, and it may be difficult for them to accept that random assignment tends to balance these out (Sawilowsky, 2004).

In the qualitative coding of the assignments, it was noted whether or not students elaborated on why random assignment was relevant to cause-and-effect conclusions (for example, talking about how confounding variables may be at play and how random assignment tends to balance these out). Again, this type of elaboration could show evidence of deeper conceptual knowledge, as it would involve interrelations between pieces of knowledge about study design and conclusions (Hiebert & Lefevre, 1986; Tennyson & Cocchiarella, 1986; Rittle-Johnson & Alibali, 1999). On these assessments, most students and groups did not explain the link between random assignment and causation without some prompt to do so. On the lab assignment, about 40% of students overall elaborated about why random assignment allows for causal claims (or conversely, why lack of random assignment makes causal claims difficult to make). On each of the two quiz contexts involving studies with random sampling but no random assignment, very few student groups (less than 8%) elaborated on why the lack of random assignment did not allow for causal claims. However, neither the lab questions nor these two quiz questions specifically asked students to elaborate on why random assignment helps with causal claims.

Nevertheless, there are other indications that students did understand the role of random assignment in making causal claims. Question #3 on the group quiz prompted more explanation on the link between random assignment and making causal claims, asking

whether confounding variables were likely to explain the observed difference in the response variable (amount of ice cream served). About 60% of student groups not only correctly answered "no" and cited the random assignment as a reason, but also provided an explanation relating to the tendency of random assignment to balance out confounding variables. In addition, during class observations, it was noted that when instructors asked students questions during their activities prompting them to elaborate on why they had answered that random assignment allowed for causal claims or why lack of random assignment did not allow for causal claims, in general, students were able to provide an explanation involving confounding variables and how random assignment should help to balance these out. This suggests that perhaps, students did have an understanding of why random assignment helps researchers make causal claims, but they did not explain this deeper understanding unless prompted to do so.

Reasoning about decision-making based on study design

Near the end of the lab assignment, students were presented with a context in which neither random sampling nor random assignment were used, and were asked whether the results from the study could be used to help a colleague decide whether to avoid peanuts during pregnancy. Almost 80% of students either mentioned the lack of ability to make causal claims, or the lack of ability to make generalizations to a defined population, making it difficult to determine whether the colleague in question was a part of this population. It was of interest to examine whether students would be more likely to bring up issues of generalization to a population that might include the colleague, or issues relating to causal claims, as previous studies found that students did not bring up relevant issues of random

assignment or experimental design when designing or critiquing a study (Derry et al., 2000; Groth, 2006).

In contrast to Derry et al.'s (2000) and Groth's (2006) findings, when asked about using results from a study to make decisions, students were more likely to mention the lack of ability to make causal claims (72%) than the lack of ability to make generalizations (43%), with many students mentioning both of these limitations. The online students did not perform as well as in-class students with this item, with 20% of them failing to mention anything about study design, and instead relying on other factors such as their own contextual knowledge. It is important to note that in this curriculum, students had just learned about both sampling issues and about experimental design, and had not learned one topic much more recently than another, as was the case in Derry et al.'s (2000) curriculum. It is also important to consider that a logical response to the final lab assignment question is that it is unwise to make decisions from only one study. However, no students gave such a response.

5.2.3 Difficulties that remain

Although students appeared to improve in their understanding of many different ideas, results from classroom observations and assessments showed that students experienced difficulty in some areas. The framework theory of conceptual change posits that students come in with initial ideas called *preconceptions*, or initial ideas about a topic, and sometimes these are incorrect (Vosniadou, 2012). This theory also posits that students can develop *misconceptions*, or erroneous interpretations of the concepts they learn, as they go through a curriculum. Some of the difficulties observed appeared to deal with incorrect preconceptions that students brought in, while others appeared to deal with misconceptions

they developed as they learned about study design and conclusions. This section discusses some major topics and concepts with which students appeared to struggle in learning about study design and conclusions.

5.2.3.1 Sample size

One potential incorrect preconception that was targeted in the activities (particularly in the "Sampling Countries" activity) was that sample size is more important than sampling method, such as believing that a sample must be large, or comprise a large part of the population in order for generalizations to be made. Some student tendencies to hold these incorrect ideas have been documented in research about students' understanding of study design and conclusions (e.g., Wagler & Wagler, 2013) and have been seen in assessment data (e.g., delMas et al., 2007; Tintle et al., 2012). In "Sampling Countries," students compared convenience samples of size 20 with random samples of size 10, and in three of the activities, students worked with relatively small sample sizes. In "Survey Incentives," students again worked with a relatively small sample size (26) when randomly sampling. During this last activity, some students were observed claiming that one could not generalize to the population due to the small sample size (despite the random sampling), and the online instructor decided to address this misconception in his activity wrap-up. However, when looking at the qualitative assessment data, there was not much evidence of students overemphasizing large sample size over method of sampling. For example, on the lab assignment, which involved convenience samples in both contexts, there were not many students who stated that one could generalize to a population due to the large sample size. There were also not many students who gave only the small sample size, without mentioning the lack of random sampling, as the reason why one could not generalize to the

population. Also, on quiz questions that involved studies with random sampling, it was rare to see student answers arguing that the sample size was too small to be generalizable. It was also rare to see answers stating that the sample size was large enough to be generalizable, without making any mention of the sampling method.

IDEA results (Appendix J) also point to evidence that students tended to begin the unit with incorrect ideas involving sample size, but then moved away from them. For example, on IDEA item 5, 17% of students taking the pretest claimed that a sample size of over 10,000 (from a population of 500,000) was too small in order to generalize, and on the posttest, the percent of students choosing this incorrect option decreased to about 3%. On IDEA item 6, students improved by almost 40 percentage points from pretest to posttest in their ability to understand that a small, random sample is preferable to a larger, biased sample. Both items 9 and 16 contained a distractor involving the small sample size, and for both of these items, the percentage of students choosing this incorrect option decreased from about one-third to under 7%.

However, one item on the IDEA test, about a study done in college dormitories, revealed incorrect thinking about the necessity of a large sample size for generalizing to a population. Item 4 (see Appendix J) involved identifying factors that do not allow results from a sample survey to be generalized to the population. This was the only item with fewer than 60% of students answering correctly on the posttest (in fact, the percent correct went from 8% on the pretest to 34% on the posttest). On both pretest and posttest, the most common response chosen was to indicate that all of the stated factors, including the sample size of 500, were a problem for generalizing. In the stem, students are told that the population is of size 5,000. Students' tendency to answer incorrectly in this manner

indicates a potential incorrect idea that a sample size must be a large portion of the population in order to make generalizations. These findings are consistent with what has been found in responses to a similar item on the CAOS test (item #38). In a large national sample of students taking CAOS (delMas et al., 2007), only about 20% of the students on the pretest, and nearly 40% on the posttest, correctly identified that the sample size of 500 was not a problem. In a different administration of CAOS, fewer than half of students identified this correct answer on pretest, posttest, and retention test, regardless of whether they took a traditional curriculum or a simulation-based curriculum (Tintle et al., 2012). However, in a previous iteration of the CATALST curriculum, over 80% of students answered a similar item on the GOALS test correctly, but this item had been modified to include only two answer options instead of four (Sabbag, 2013).

It should be noted, however, that the item about college dormitories appeared differently on the CAOS, GOALS, and IDEA assessments. On the CAOS test and IDEA, students are asked to choose which option does NOT affect a college official's ability to generalize the survey results to all dormitory students. On CAOS, option A (the correct option) reads "Five thousand students live in dormitories on campus. A random sample of only 500 were sent the survey." On IDEA, based on feedback given by one of the external reviewers, the population size was moved to the stem, and option A reads: "Only 500 students were sent the survey." On the GOALS test used by Sabbag (2013), students are simply asked to agree or disagree with the statement "The survey results cannot be generalized to the population of students currently living in dormitories because it was sent to only 500 students." Although students' tendency to get this item incorrect may point to incorrect ideas regarding sample size, it is also possible that the negative wording of the

question may have caused problems with cognitive load (although the word "not" was bolded and capitalized), and rewriting the item as it was done in GOALS may be better able to capture the incorrect notion that a sample must be large relative to the population in order to be generalizable.

Another incorrect preconception targeted in the activities is that sample sizes in two groups must be equal in order to make inferences, such as the tendency to believe that all methods of assignment to groups are appropriate as long as there are equal groups (Wagler & Wagler, 2013). This was targeted by having students randomly assign an odd number of subjects to groups in the "Survey Incentives" activity. Class observation notes indicated that the unequal sample sizes were problematic for some students. For example, some online students suggested nonsensical ways of assigning the 25 subjects to the two groups (incentive and control), such as creating 5 groups of 5 subjects each. However, on the IDEA test, the incorrect notion that sample sizes must be equal did not appear very prevalent. On item 17 (see Appendix J) only about 15% of students on the pretest, and about 9% on the posttest, chose distractor C which said that one could not draw valid conclusions because the sample sizes were not the same. This may be because the students in the class had previously performed randomization tests to compare groups with unequal sample sizes.

5.2.3.2 Terminology

Throughout the activity observations and analysis of the assignments, there was evidence of students struggling with correct use of terminology. This included difficulty recognizing generalization statements and causal statements, problems understanding the terms "generalization" and "causation," and using colloquial meanings of the terms "random" and "bias" instead of their statistical definitions.

During the activities, students struggled with some basic definitions that had been introduced in readings that were to be completed either prior to class or during activities. At the beginning of the first activity, "Sampling Countries," students were presented with a brief reading defining the terms "parameter" and "statistic," but according to class observation notes, many students had trouble with questions involving these terms. Before the "Strength Shoe" activity, students were required to complete a reading called "Establishing Causation" which introduced the terms "explanatory" and "response" variables. (Prior to this unit, students had learned about "treatment" and "response" variables using experimental data in other activities.) During the "Murderous Nurse" activity, students were observed showing confusion about what an explanatory variable was, despite the definition having been in the reading that had been assigned.

The instructors of the CATALST course had shared prior to the study design unit that students typically began activities with little, if any, large group introduction. The activity lesson plans were therefore designed so that large group discussion typically happened as wrap-up, as was customary in the course, without much discussion happening before the activities. However, the problems that students had identifying basic terms suggest that perhaps some prior discussion to define key terms, as well as providing motivation for students to complete the readings, would be beneficial. It is unclear what percentage of students actually completed the required readings before class. The instructor of sections 1 and 3 gave a pop quiz on the "Establishing Causation" reading, but collection of these responses was not a part of the proposed data collection for this project.

In addition, there was evidence that some students had difficulty with understanding what it meant to make a "generalization" or a "causal claim." For example,

during the "Murderous Nurse" online activity discussion, when asked about being able to generalize the results to the population of shifts at the hospital, many students referenced confounding variables that would make Gilbert's shifts different from other shifts, thus referring to the lack of ability to make causal claims, not generalizations. On the IDEA pretest, almost all students were correctly able to distinguish between headlines that made association-only claims and headlines that made causal claims (items 12-15). However, on the group quiz, about 12% of student groups gave answers that suggested they had failed to recognize whether or not a headline was making a generalization, and about 14% of groups gave answers suggesting that they had failed to recognize whether or not a headline was making causal claims. During the quiz, in-class groups were observed debating what types of claims a headline made. For example, in question 1, some students were observed arguing whether or not the headline "In U.S., Moderate Drinkers Have Edge in Emotional Health" was making a causal claim. These difficulties suggest that it would be beneficial to incorporate into activities and large group discussion more practice with talking about what it means to make a generalization, what it means to make a causal claim, and what these claims can look like.

Problems with terminology can also stem from students' prior experience with colloquial meanings of terms used in statistics. As suggested by Vosniadou's (2012) framework theory, students come into a course with preconceptions, or initial ideas (which are not necessarily incorrect). In this case, students came in with their own definitions of "random" and "bias" as suggested by colloquial language. For example, students have been known to think of the colloquial definition of "random" as "by chance," "haphazard," "unexpected," or "without a pattern" (Kaplan et al., 2009; 2014; Smith & Hjalmarson,

2013). There was some evidence of students using these colloquial definitions in the classroom observations. For example, some students stated that when they tried to choose a sample of countries that was representative of the countries in the world, they chose them "randomly" or chose "random countries." Also, in large group discussion, when students were asked about bias, one student responded that it was important to think about who is collecting the data, referring to a researcher possibly being "biased," rather than referring to the sampling bias that was being discussed. On written assignments, use of these colloquial definitions of "random" and "bias" was generally not seen, but it was seen in the classroom observations.

Distinguishing between random assignment and randomization testing

The CATALST course in which the study design curriculum was taught focuses on teaching students the logic behind inference using simulation and randomization-based methods. Students had performed randomization tests many times prior to the study design unit, and had also learned about computing bootstrap intervals. The activities used in this unit relied on notions of repeated sampling to teach about sampling bias and repeated random assignments to teach about balancing confounding variables. The advantage of having the study design unit happen so late in the semester is that students already had experience with repeated sampling by learning about bootstrapping. Students also had experience with random allocation in the randomization tests they had conducted.

However, one difficulty that occurred during the "Murderous Nurse" activity is that some students were observed stating that one could make causal claims from this study due to the random assignment, when they were in fact referring to the random re-allocation done in the randomization test procedure. Building conceptual knowledge involves tying

together pieces of information (Rittle-Johnson & Alibali, 1999; Tennyson & Cocchiarella, 1986). Students could use their previous knowledge about repeated sampling and about random re-allocation, but this could also result in an improper interpretation of the concepts they were learning, or the building of *misconceptions* (Vosniadou, 2012). Confusing random assignment from an original study with random re-allocation from a randomization test procedure is an example of a possible misconception developed by the students.

5.2.3.3 Disbelief in the effectiveness of random assignment

One finding from previous research is that students tend to have difficulty believing that random assignment can balance out confounding variables, but having students explore multiple confounding variables for many random assignments can alleviate this difficulty (Sawilowsky, 2004). The activities in this study design curriculum were designed to have students examine multiple confounding variables when comparing groups after many random assignments. During the "Strength Shoe" activity, some students were observed expressing disbelief in the effectiveness of random assignment to help enable causal claims between the type of shoe and observed differences in jumping ability. For example, some students stated that the genetic "X-factor" could still differ between the groups after the random assignment was conducted. Students' hesitation to make causal claims may partially stem from a healthy skepticism in the ability to use a single study to make decisions, or from a valid concern that a "bad" random assignment could still result in imbalance between groups. However, these types of hesitations were not predominantly heard, suggesting that examining multiple confounding variables for many random assignments may alleviate this disbelief in the usefulness of random assignment (Sawilowsky, 2004).

In contrast to the skepticism observed during the "Strength Shoe" activity, some students were too quick to make causal claims in the "Murderous Nurse" activity. Some students used the extremely low *p*-value to justify that the nurse was indeed killing patients, despite the fact that this was an observational study. On the last question of the lab assignment, a few students also used a low *p*-value from an observational study to justify the avoidance of peanuts during pregnancy to prevent allergies. Therefore, while students were at times overly cautious about making causal claims based on experiments, at other times they were too quick to make causal claims based on observational studies. However, this was not the case for the majority of students. (Less than 5% of students on the lab assignment used the low *p*-value to make causal claims.) It is unclear why students were at times quick to make causal claims based on observational studies, but one possible explanation is that students tend to focus on their own contextual knowledge to reason about statistical questions (Wroughton et al., 2013). The "Murderous Nurse" activity is based on a real study in which a nurse was convicted of murder, and students were encouraged to look up information about this story online after finishing the activity. It is possible that students' knowledge that the nurse was actually convicted could have influenced their hastiness to make causal claims.

Some amount of skepticism about the ability to make a causal claim from a randomized experiment was seen on the group quiz. On this quiz, the second context presented to students involved a randomized experiment to examine whether bowl size affected amount of ice cream served, and it was found that those with larger bowls tended to eat significantly more ice cream. Question 3 asked students if it was likely that other factors explained this difference. Very few student groups gave answers both

acknowledging that random assignment had been done, and still arguing that confounding variables could explain the observed difference. Just under one-fifth of student groups gave possible confounding variables and stated how they could affect amount of ice cream served, without mentioning that random assignment had been done. For these one-fifth of groups, it is unclear whether they missed the fact that random assignment was done in the stem of the question, or whether they believed that despite this random assignment, confounding variables were still likely to affect results.

In summary, although the hesitation to make causal claims based on a randomized experiment with statistically significant results documented by Sawilowsky (2004) was observed on occasion, it was not the case for the great majority of students. Most students appeared to successfully recognize the role of random assignment in helping to enable causal claims.

5.2.4 Distinguishing between random sampling and random assignment

As summarized above, the IDEA posttest, group quiz, and lab assignment provide evidence that many students were able to demonstrate correct reasoning about sampling and its implications for generalization, and about assignment to groups and its implications for making causal claims. However, it was also of interest in this study whether students would successfully be able to distinguish between the different roles of random sampling and random assignment, or whether they would show confusion between the two, as has been found previously among introductory statistics students (e.g., Derry et al., 2000). While this confusion was not prevalent overall, it definitely was present among a considerable portion of student responses to some items on both IDEA and on the open-ended assessments.

Developing incorrect interpretations after being exposed to course content is an example of what Vosniadou (2012) might call a *misconception* in the framework theory of conceptual change. One misconception that was anticipated in this curriculum involved difficulty distinguishing between the purposes of random sampling and random assignment. Evidence of this misconception was observed somewhat, but was not prevalent, in student discussion during class activities, even during the last activity in the curriculum. For example, during the "Survey Incentives" activity, some online students suggested that random assignment allowed one to generalize to a population *and* make causal claims. In particular, section 2 students revealed a lack of ability to distinguish between random sampling and random assignment in large group discussion, resulting in extra time being spent by the instructor to clarify these issues.

On the IDEA test, there were four items that involved distractors indicating possible confusion between random sampling and random assignment. For only one of these items (item 11) did fewer than 10% of students choose the distractor indicating this potential confusion. On this item, students were asked to identify the best study design for being able to conclude that taking a vitamin causes a change in cholesterol level. Fewer than 2% of students chose the incorrect option stating that a survey should be sent to a random sample of patients, and the vast majority correctly identified that an experiment with random assignment to take vitamins or a placebo would be the best design.

However, there were three IDEA items (items 9, 16, and 18) for which more than 10% of students chose answer options that could indicate confusion between random sampling and random assignment. On item 9, 22% of students on the posttest chose incorrect option D, identifying random assignment as the reason for why a headline that

made a generalization was problematic (see Appendix J). However, it was noted during observations of the group quiz that some students had problems identifying whether statements made generalizations or causal claims. Therefore, it is unclear whether these students chose option D on item 9 because they thought random assignment was necessary to make a generalization, or because they interpreted the headline incorrectly as making a causal claim. It is also interesting to note that the percent of students who chose incorrect option D increased from 9% to 22% from pretest to posttest. If students choose this answer option because they think that random assignment is necessary for making a generalization, then these results show a possible increase in confusion between random sampling and random assignment for a noticeable minority of students.

Item #16, which involved identifying that a strong, statistically significant correlation does not imply causation, and item #18 about identifying the purpose of random assignment, were two of the items on IDEA with the most improvement. However, both items had a considerable number of students choose incorrect answer options that showed potential confusion between random sampling and random assignment. On item #16, just over 10% of all students on the posttest chose the incorrect answer option D that causation could be inferred due to the fact that a random sample was used. On item #18, about 15% of all students on the posttest incorrectly chose option C stating that the purpose of random assignment is to ensure that participants are representative of the larger population. However, only 5% of students chose *both* of these incorrect answer options, suggesting that a true confusion between random sampling and random assignment may not be as prevalent as it would appear from examining each item individually. For both items, the percentage of students who chose that incorrect answer option decreased from pretest to

posttest, showing that they were less likely to display these incorrect ideas after the study design curriculum than before.

On the lab assignment and group quizzes, misunderstandings about which study design helps with which type of conclusion were noticeably present, but did not represent the majority of responses. It was rare to see answers indicating that both random sampling and random assignment were needed to generalize, or to make causal claims. It was more common to confuse the purpose of random sampling with the purpose of random assignment, and vice-versa. This type of confusion is consistent with behavior found in the study by Derry et al. (2000).

For example, on the lab assignment, about 24% of students displayed at least one misunderstanding related to which study design helps with which type of conclusion. The most common misunderstanding, displayed by just over 10% of students, was to bring up only random assignment (or lack thereof) when the question was asking about generalization, but not causation. For example, a student might state that due to the random assignment, one could generalize to the population. Similarly, on the ice cream context (questions 3-4) on the group quiz, about 25% of students displayed at least one misunderstanding about which study design helps with which type of conclusion. On the questions related to this context, the most common misunderstanding related to types of conclusions displayed was bringing up the lack of random sampling when asked whether it was likely that factors other than bowl size could explain the differences between groups in amount of ice cream served. Some student groups cited the lack of random sampling, or the fact that the sample was a convenience sample of nutritionists from Massachusetts, rather than recognizing that the random *assignment* was relevant to this question.

In fact, most students did not display confusion between random sampling and random assignment on open-ended assessments, and there was some evidence that they could recognize and correct this type of confusion. For example, on the lab assignment, question 13 was designed specifically to try to diagnose confusion between random sampling and random assignment, but over 80% of students correctly identified the misconception and corrected it with an appropriate explanation. (The online students, however, did not perform as well as the in-class students with this question.)

It was anticipated that some students might give responses that made it ambiguous whether they understood the distinction between random sampling and random assignment. For example, a statement like "we cannot make causal claims or generalizations because there was no random sampling or random assignment" is true, but does not make it clear which study design corresponds to which type of conclusion. Also, a response that talked about "randomness" being needed for generalization or for causation made it ambiguous as to whether the randomness needed to happen in the sampling, or the assignment. However, ambiguous student answers were not common. It was more common for students to give extraneous information in their answers, such as talking about both generalization and causation when only asked about one or the other. This type of behavior was quite common on the quiz, and also was displayed by about one-quarter of students on the lab. Giving extraneous information does not mean that students are confusing generalization with causation, but it's possible that students were unclear about whether some questions asked about generalization or causation, and thus chose to address both in their answer. Another explanation for giving extraneous information is that students connected both understandings of generalization and causation, so a question asking about

generalization also prompted them to think about causation. If being asked about one aspect of scope of inferences (such as generalization) prompted students to additionally think about another type of inference (such as causation), this could suggest there is some evidence of deeper learning and connection between concepts related to scope of inference (Hiebert & Lefevre, 1986; Rittle-Johnson & Alibali, 1999; Tennyson & Cocchiarella, 1986).

In summary, confusion between random sampling and random assignment was not highly prevalent on assessment answers. However, this confusion was still present at the end of the curriculum for a small, but noticeable portion of students. Even though students went through a curriculum designed to help them to distinguish between the purposes of random sampling and random assignment, this distinction can still be challenging for students to learn, and "pervasive confusion" (Derry et al., 2000) between the two may be difficult to eliminate for some students, even after experiencing a curriculum designed to help students understand the distinction.

5.3 Study limitations

While this curriculum was designed to teach students about the usefulness of random sampling and random assignment in statistical studies, neither random sampling nor random assignment were used in this particular study. Although this study can provide useful insight into students' understanding of study design and conclusions, there are limitations to any generalizations and causal claims that can be made.

The students who participated in this curriculum were enrolled in an introductory statistics course that fulfills a mathematical thinking general education requirement at the University of Minnesota. Although no demographic information was collected on the

students, the course historically has tended to draw liberal arts students who are not majoring in mathematics, statistics, or physical sciences (Garfield et al., 2012). During class observations, it was also noted that females were represented at a higher rate than males. Therefore, the results from this study may not be generalizable to all students who take an introductory statistics course. Moreover, the instructors who implemented this curriculum are likely not representative of all introductory statistics instructors who might teach this curriculum. The instructors were all highly experienced teachers of statistics, familiar with active learning pedagogy and statistics education research, and had taught this particular introductory course prior to this study. This means that instructor expertise, and not just the curriculum, could be a contributing factor to observed improvement in students' test performance on the IDEA instrument.

Although there were improvements on IDEA performance from pretest to posttest, no causal claims about this curriculum can be made. Because this introductory statistics course is taught the same way across all sections of the course each semester, the study design unit was implemented in all sections. It was not possible to make comparisons with another curriculum, and no data were gathered about students' understanding of study design and conclusions in prior semesters of the course. Therefore, it is impossible to tell whether this particular study design unit improves students' understanding any more or any less than another curriculum.

Although students' IDEA responses provided valuable insight into their understanding of specific topics before and after the study design unit, there are limitations to using this instrument. Even though the IDEA instrument was developed using guidelines for assessment development (AERA, APA, & NCME, 1999) such as beginning with a test

blueprint and soliciting feedback from expert reviewers, the instrument had relatively low reliability as measured by coefficient omega (McDonald, 1999). Students took the assessment online outside of class, and the only requirements were to take the pretest before the start of the curriculum and the posttest within a few days after the curriculum. Measurement error may have been introduced by factors such as student guessing and environmental variability resulting from distractions and differences in testing locations. Students may also have varied in how long they waited after the curriculum to take the posttest, and their responses may have been affected by how recently they had experienced the final activity of the unit before they took the test. For ease in assigning grades, instructors required that students complete IDEA as part of their lab assignments, but did not award more points to students who got more correct answers on the posttest. Therefore, it is difficult to know how motivated students were to do their best. Also, some items had very high performance on both pretest and posttest, and made it difficult to measure any changes in student understanding related to those learning goals.

Due to the structure of the introductory statistics course, the study design curriculum was implemented about two-thirds of the way through the semester. Students already had experience conducting bootstrap intervals and conducting randomization tests, and had already worked with real data from experimental studies. Although they had not explicitly learned yet about study design and conclusions, they had considerable prior knowledge of statistical topics related to study design. Therefore, the performance on the pretest for this group of students likely does not accurately represent typical prior knowledge about study design concepts for students who have not yet entered into a statistics course. Moreover, there were nine items (out of 22 total) on the IDEA instrument

that more than 80% of students answered correctly on the pretest and posttest. Thus, many items on the IDEA test may not have been able to differentiate well between students with higher levels and lower levels of understanding.

Another limitation of this study is that the IDEA test was only administered twice: Once immediately prior to the unit and once immediately after the unit. Although it was of interest to measure students' retention of concepts related to study design and conclusions at the very end of the course, a third administration was not possible. This was because another graduate student conducting her own dissertation research needed to administer an assessment at the end of the course, and it was not desirable for students to be over-tested. Although there is information about how students' performance on IDEA changed from just before to just after the study design unit, it is unknown how much of their knowledge was retained.

Limitations also arise when considering the formats in which the unit was offered, and the varying instructors. Because three sections were in-class and one section was online, and because there was instructor variation in style of teaching, the students did not all experience the curriculum exactly the same way. The in-class format allowed for more back-and-forth discussion among students and between students and instructor, but not all discussions could be heard by the researcher or co-observer, nor could the camera record all groups' discussions. Online, the discussions were recorded on the discussion boards and all student discussion could be seen, but the format only allowed for discussion of some key questions and due to the asynchronous nature of the class, there was less time for back-and-forth interaction. This meant that if some students in a group answered incorrectly or misinterpreted a question, they could lead the rest of the group down the wrong path until

the instructor was able to intervene. In-class instructors could lead large group wrap-up discussions, whereas the online instructor could not lead an additional wrap-up discussion. Instead, the online instructor posted a video or written paragraphs and gave feedback to students' group summaries.

Nevertheless, although there were differences in how each section experienced the curriculum, analysis of the IDEA scores did not reveal striking differences between sections. Only sections 1 and 4 were significantly different from each other on the pretest, but no significant differences existed among sections on the posttest. When performance on IDEA items was examined individually (Appendix J), there were sometimes observable differences between sections, but there were no sections that did consistently better or worse than any others across the different items. These findings suggest that the study design curriculum may be robust to variations in instructions and delivery of the curriculum.

5.4 Implications for teaching

According to statistics education recommendations for an introductory statistics course (GAISE, 2016), students should understand why random sampling facilitates generalizations to the population from which the sample was gathered, and why random assignment in experimental design facilitates cause-and-effect conclusions. This involves thinking critically about research as educated citizens, being able to identify sources of bias in sampling and reasoning about what conclusions can be made from observational and experimental studies (Utts, 2003). Although topics of study design and conclusions are arguably important in an introductory course, they are also difficult for students to understand, as they involve the integration of many different concepts.

Some changes to the curriculum were suggested by the instructors and observers of the course, in particular with regards to teaching about sampling and bias. In the "Sampling Countries" and "Survey Incentives" activities, students experienced repeated sampling and observed how statistics obtained from repeated random samples were centered at the population parameter. In discussions with the instructors of the course and with an observer, it was brought up that the activity and wrap-up focused on the *consequence* of random sampling (that the sample means tend to be at the parameter, on average) but did not emphasize *why* this happens: In simple random sampling, each unit is equally likely to be selected. Also, after the "Sampling Countries" activity, instructors shared feedback that although there was a good discussion of bias, there had not been enough discussion on what it means to generalize to a population. Although there was a paragraph at the end of the "Sampling Countries" activity that explained generalization, students did not get practice identifying statements that generalized to populations in different contexts. This is why the instructors decided to add additional wrap-up addressing generalization on the second day of the curriculum. In the future, adding more exploration of why random sampling is unbiased, and what it means to generalize to a population, may help students develop a fuller conceptual understanding.

Although past research can be helpful in pointing out potential difficulties in student understanding (e.g., Derry et al., 2000; Groth, 2006; Sawilowsky, 2004; Wagler & Wagler, 2013), sometimes, issues arose that were not anticipated by the researcher. One of the major unanticipated issues for students in this study was the terminology. Students might encounter definitions of basic terms in readings prior to activities (e.g., "explanatory variable," "parameter," "statistic"), but were unsuccessful in identifying these definitions

correctly when asked about them in activities. This may point to a need to incentivize students to make sure they complete assigned readings ahead of time (e.g., pop quizzes), or a need to briefly go over basic terms in an introductory discussion prior to activities. Having students construct a glossary of terms could also be helpful in aiding their memory of these terms throughout the course.

While students had some difficulty remembering basic definitions, sometimes the terminology was confusing for them at a deeper level. For example, the terms "random," "bias," and "confounding" can relate to both topics of sampling and generalization, and assignment and causation. This could make it difficult for students to distinguish between the purposes of random sampling and random assignment. For example, the term "random" is common to both of these study design methods, and this may contribute to difficulties distinguishing between the role of randomness in sampling and the role of randomness in assignment to groups. The term "bias" can be used to refer to a sampling method that tends to yield a sample that is not representative of the population, or to "researcher bias" in assigning groups that are not approximately equivalent in characteristics. In the design of this curriculum, the term "confounding" was used only to refer to variables that could explain an association between explanatory and response variables. However, students also began to use the phrase "confounding variables" to refer to variables that could make the sample different from the population. The use of common terms to describe concepts related to both generalization and causation can help students tie together concepts, but it is important for students to make a clear distinction between how these terms are used to describe the purpose of random sampling and how they are used to describe the purpose of random assignment.

The findings from this study imply that concepts of study design and conclusion are not merely viewed as isolated topics, but relate to other topics they see in the curriculum and should be integrated with their knowledge of other concepts. For example, students had concerns about sample size, such as indicating that a relatively small sample size could not provide generalizable results, even if random sampling was used. Also, students expressed concern with being able to make causal claims when the groups being assigned to treatments were small in size. Although sample size does not directly affect bias, students' concerns about sample size are still valid and should be addressed. For example, the online instructor pointed out in the discussion boards that sample size is accounted for in statistical inference methods, such as affecting the width of confidence intervals and the size of *p*-values. Students should also visit topics of sample size and variability, and how these can relate to issues of estimation and inference. For instance, random assignment will more evenly balance out groups for larger sample sizes than for smaller ones. Also, while random sampling helps to eliminate bias, another issue in interval estimation is sampling variability, which is smaller for larger samples. Additionally, in a randomization-based curriculum, it is important for students to distinguish between random assignment in the original study and random re-allocation in a randomization test, so that they can perform statistical inference and then use the results to make conclusions that are also supported by the design of the study.

As suggested by cognitive researchers (e.g., Vosniadou & Brewer, 1987; Vosniadou, 2013), learning concepts and conceptual change takes time. The findings from this study suggest it is not feasible to expect students to completely master conceptual knowledge of the purposes of random sampling and random assignment after a single brief

unit. Instead, it may be advantageous to revisit these topics throughout the course and in new contexts. Moreover, the use of studies to make conclusions and decisions is more complex than it may first appear to students. In real life, many studies they will encounter do not use random sampling or random assignment, but students should learn that these studies can still contribute valuable research despite their limitations. Students have valid intuitions that a single study is not enough to make decisions, and this concern should be acknowledged. Instruction should continue to take into account the uncertainty that is inherent to statistics, and care should be taken so that instruction does not imply that one can make absolute generalization statements from studies with random samples, or irrefutable causal claims from studies with random assignment. Students should learn the purposes behind the use of random sampling and random assignment, but also understand that the real world is complex and these study designs may be difficult to achieve.

5.5 Implications for research

This was an exploratory study examining students' understanding of concepts related to study design and conclusions, and has various limitations that warrant future research. For example, while there was a significant increase in performance from pretest to posttest on IDEA, the reliability measures were not as high as desired. This points to a need for an assessment that can more reliably measure conceptual understanding of random sampling and generalization, and random assignment and causation. Also, because it was not feasible to administer IDEA at the end of the course, it is unknown how much of students' knowledge about study design and conclusions was retained at the end of the semester. Future research could also measure how much knowledge is retained even after students have spent time away from statistics after completion of the course.

In this study, it was not possible to compare different ways of teaching study design, but future research could explore different curriculum variations. For example, the topic of sampling was placed before random assignment, because in the data collection process, a sample is gathered before any potential group comparisons are made. It is unclear whether or not it makes a difference to teach random sampling first, or random assignment first. Also, most of the textbooks reviewed prior to the design of this study (e.g., Agresti & Franklin, 2009; Lock et al., 2013; Moore, 2010) place topics of random sampling and random assignment in close proximity to each other, either within the same book section or in consecutive sections. Future research could also explore whether it makes a difference to teach topics of sampling and topics of assignment to groups consecutively, or to have more separation between them as was done in the curriculum by Derry et al. (2000). Future research could also explore whether different placements of these topics in the curriculum affects student learning. For example, some textbooks teach study design at the very beginning, possibly following the logic that data collection is the first step of conducting statistical studies (e.g., Devore & Peck, 2005; Lock et al., 2013), while others teach descriptive statistics and exploring relationships before mentioning data collection (e.g., Agresti & Franklin, 2009; DeVeaux et al., 2009). Whatever the placement in the curriculum, it is important to consider students' prior knowledge before studying topics of study design. For example, if students have prior experience with randomization testing, it is important for them to learn to distinguish between random assignment in study design and random re-allocation in a randomization test.

Cognitive research suggests that conceptual change is slow and gradual (e.g., Vosniadou, 2012), and thus it would be of interest to examine how students' understanding

of study design and conclusions changes throughout a course (or even a sequence of courses, if applicable) where students revisit these topics. Future research may also consider the impact of starting to teach topics of study design earlier in students' education, as has been recommended by the Common Core State Standards for Mathematics (http://www.corestandards.org/Math/). Generalization to a population and making causal claims are not isolated topics, but topics that apply when making conclusions from any statistical study. Therefore, they can be revisited many times, and future research could explore how students' understanding of these concepts evolves as they continue to revisit study design issues in different contexts.

5.6 Conclusion

This study explored students' understanding of study design and conclusions by introducing a study design unit with activities and assessments in an undergraduate introductory statistics course. Although this was not an experimental study with random assignment to treatments, there is some evidence that the curriculum may have had a positive effect on student learning. For example, although the reliability of the IDEA instrument was lower than desired, results showed a significant increase in overall scores from pretest to posttest. There was also a significant increase in performance for various items, and all items except one had more than 60% of students answering correctly on the posttest. Moreover, there was noticeably higher performance on some IDEA items on the posttest in this study, when compared to performance on similar items in previous studies. Classroom observations and qualitative analysis of a group quiz and lab assignment revealed that most students were able to successfully make connections between random sampling and generalization, and between random assignment and causation, although

confusion distinguishing between random sampling and random assignment persisted for a small, but noticeable amount of students. This suggests that although the curriculum may have helped student learning, it is unrealistic to expect introductory statistics students to completely master conceptual knowledge of study design and conclusions after a short unit.

This study design unit was implemented in all four sections of an introductory statistics course, and it was not compared to any other ways of teaching study design. However, the activities and assessments were designed based on reviews of literature on conceptual change and of the limited statistics education literature on students' understanding of random sampling and random assignment. Therefore, the activities and assessments used in this study may be valuable for instructors looking for new ways to teach about study design and conclusions in their introductory statistics courses.

References

Abramovich, S., & Ehrlich, A. (2007). Computer as a medium for overcoming misconceptions in solving inequalities. *Journal of Computers in Mathematics and Science Teaching*, *26*(3), 181–196.

Agresti, A., & Franklin, C. (2009). *Statistics: The art and science of learning from data* (2nd ed.). New Jersey: Pearson Prentice Hall.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (1999). *Standards for educational and psychological testing.* American Educational Research Association.

American Statistical Association. (2005). *Guidelines for assessment and instruction in statistics education: College report.* Alexandria, VA. http://www.amstat.org/education/gaise/

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369.

Ayer, A. J. (1965). Chance. *Scientific American*, *213*, 44–54.

Ayton, P., Hunt, A. J., & Wright, G. (1989). Psychological conceptions of randomness. *Journal of Behavioral Decision Making*, *2*(4), 221–238.

Bar-Hillel, M., & Wagenaar, W. A. (1991). The perception of randomness. *Advances in Applied Mathematics*, *12*(4), 428–454. doi:10.1016/0196-8858(91)90029-I

Baroody, A. J., Feil, Y., & Johnson, A. R. (2007). Research commentary: an alternative reconceptualization of procedural and conceptual knowledge. *Journal for Research in Mathematics Education, 38*(2), 115–131.

Beckman, M. D., delMas, R. C., and Garfield, J. (2017). Cognitive transfer outcomes for a simulation-based introductory statistics curriculum. *Statistics Education Research Journal, 16*(2), 419-440.

Bellhouse, D.R. (1988). A brief history of random sampling methods. In P.R. Krishnaiah & C.R. Rao (Eds.), *Handbook of Statistics 6: Sampling* (pp. 1-14). New York: Elsevier.

Biehler, R., Ben-Zvi, D., Bakker, A., & Makar, K. (2013). Technology for enhancing statistical reasoning at the school level. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, and F. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 643–690). New York: Springer.

Bransford , J. D. , Brown , A. L. , & Cocking , R. R. (1999). *How people learn: Mind, brain, experience, and school* . Washington , DC : National Research Council .

Chance, B., Ben-Zvi, D., Garfield, J., & Medina, E. (2007). The role of technology in improving student learning of statistics. *Technology Innovations in Statistics Education*, *1*(1).

Charles, Eric P. (2005). "The correction for attenuation due to measurement error: clarifying concepts and creating confidence sets." *Psychological methods*, *10*(2), 206-226.

Cobb, G. W. (1998). *Design and analysis of experiments*. New York: Springer.

Cornfield, J. (1959). Principles of research. *American Journal of Mental Deficiency, 64*(2), 240-252.

Dean, A., & Voss, D. (1999). *Design and analysis of experiments*. New York: Springer.

Dega, B. G., Kriek, J., & Mogese, T. F. (2013). Students' conceptual change in electricity and magnetism using simulations: A comparison of cognitive perturbation and cognitive conflict. *Journal of Research in Science Teaching*, *50*(6), 677–698. doi:10.1002/tea.21096

delMas, R., Garfield, J., & Chance, B. (2004, April). Using assessment to study the development of students' reasoning about sampling distributions. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA. Retrieved from http://www.gen.umn.edu/faculty_staff/delmas/AERA_2004_samp_dist.pdf.

delMas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, *6*(2), 28–58.

Derry, S. J., Levin, J. R., Osana, H. P., Jones, M. S., & Peterson, M. (2000). Fostering Students' Statistical and Scientific Thinking: Lessons Learned from an Innovative College Course. *American Educational Research Journal*, *37*(3), 747–773. doi:10.3102/00028312037003747

DeVeaux, R., Velleman, P., & Bock, D. E. (2009). *Intro Stats* (3rd ed.). Boston, MA: Pearson Education, Inc.

Devore, J., & Peck, R. (2005). *Statistics: The Exploration and Analysis of Data* (5th ed.). Belmont, CA: Thomson Brooks/Cole.

Dietz, E. J. (1993). A cooperative learning activity on methods of selecting a sample. *The American Statistician*, *47*(2), 104–108.

Donovan, M. S., & Bransford, J. D. (Eds.). (2005). *How students learn: History, science and mathematics in the classroom.* Washington, DC: National Academies Press.

Enders, C. K., Laurenceau, J.P., & Stuetzle, R. (2006). Teaching Random Assignment: A Classroom Demonstration Using a Deck of Playing Cards. *Teaching of Psychology*, *33*(4), 239–242. doi:10.1207/s15328023top3304_5

Falk, R. (1991). Randomness—an ill-defined but much needed concept. *Journal of Behavioral Decision Making, 4*(3), 215-218.

Falk, R., & Konold, C. (1994). Random Means Hard to Digest. *Focus on Learning Problems in Mathematics*, *16*(1), 2–12.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, *104*(2), 301–318. doi:10.1037/0033-295X.104.2.301

Fisher, R.A. (1925). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain,* 33: 503-513.

Ford, J. (1983). How random is a coin toss? *Physics Today*, *36*(4), 40–47.

GAISE College Report ASA Revision Committee, "Guidelines for Assessment and Instruction in Statistics Education College Report 2016," http://www.amstat.org/education/gaise.

Garfield, J., & Ben-Zvi, D. (2008). *Developing students' statistical reasoning: Connecting research and teaching.* New York: Springer.

Garfield, J., & delMas, R. (2010). A web site that provides resources for assessing students' statistical Literacy, reasoning and thinking. *Teaching Statistics*, *32*(1), 2–7. doi:10.1111/j.1467-9639.2009.00373.x

Garfield, J., delMas, R., & Chance, B. (2002). The Assessment Resource Tools for Improving Statistical Thinking (ARTIST) Project. NSF CCLI grant ASA-0206571. Retrieved from https://apps3.cehd.umn.edu/artist/index.html

Garfield, J., delMas, R., & Zieffler, A. (2012). Developing statistical modelers and thinkers in an introductory, tertiary-level statistics course. *ZDM*, *44*(7), 883–898.

Groth, R. E. (2006). An exploration of students' statistical thinking. *Teaching Statistics*, *28*(1), 17–21. doi:10.1111/j.1467-9639.2006.00003.x

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334.

Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: an introductory analysis. In J. Hiebert (Ed.), *Conceptual and Procedural*

Knowledge: The Case of Mathematics (pp. 199–223). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396). doi:10.1080/01621459.1986.10478354

Kac, M. (1983). Marginalia: What is random? *American Scientist*, *71*(4), 405–406.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430–454. doi:10.1016/0010-0285(72)90016-3

Kaplan, J. J., Fisher, D. G., & Rogness, N. T. (2009). Lexical ambiguity in statistics: What do students know about the words association, average, confidence, random and spread? *Journal of Statistics Education*, *17*(3), 1–19.

Kaplan, J., Rogness, N. T., & Fisher, D. G. (2014). Exploiting lexical ambiguity to help students understand the meaning of random. *Statistics Education Research Journal*, *13*(1), 19–24.

Kempthorne, O. (1977). Why randomize? *Journal of Statistical Planning and Inference*, *1*(1), 1–25.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction*, *6*(1), 59–98.

Kruskal, W., & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review/Revue Internationale de Statistique*, *48*(2), 169–195. doi:10.2307/1403151

Kruskal, W. H., & Mosteller, F. (1988). Representative sampling. *Encyclopedia of Statistical Sciences*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/0471667196.ess2253.pub2/full

Labov, J. B., & Firmage, D. H. (1994). Introducing concepts of random ordering and random assignment of subjects: computer-assisted classroom & laboratory exercises. *The American Biology Teacher*, *56*(3), 169–173. doi:10.2307/4449782

Levy, P., & Lemeshow, S. (1999). *Sampling of populations: Methods and applications*. New York: John Wiley & Sons.

Lock, R., Lock, P. F., Lock Morgan, K., Lock, E. F., & Lock, D. F. (2013). *Statistics: Unlocking the power of data*. Hoboken, NJ: John Wiley & Sons.

Lohr, S. (2010). *Sampling: design and analysis* (2nd ed.). Boston: Cengage Learning.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction*, *16*(3), 285–265. doi:10.1207/s1532690xci1603_3

Moore, D. S. (1997). New pedagogy and new content: The case of statistics. *International Statistical Review*, *65*(2), 123–137.

Moore, D. S. (2001). *Statistics: Concepts and Controversies* (5th ed.). New York: W.H. Freeman and Company.

Moore, D. S. (2010). *The Basic Practice of Statistics* (5th ed.). New York: W.H. Freeman and Company.

Moore, D. S., & McCabe, G. P. (1999). *Introduction to the Practice of Statistics* (3rd ed.). New York: W.H. Freeman and Company.

Nekvasil, N. & Liu, D. (2016). *Gallup.* "In U.S., moderate drinkers have edge in emotional health." Retrieved from: http://www.gallup.com/poll/188816/moderate-drinkers-edge-emotional-health.aspx?g_source=CATEGORY_WELLBEING&g_medium=topic&g_campaign=tiles

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 558–625.

Olive, J., & Makar, K. (2010). Mathematical knowledge and practices resulting from access to digital technologies. In C. Hoyles & J.B. Lagrange (Eds.), *Mathematics education and technology revisited: Rethinking the terrain* (pp. 133–177). New York, NY: Springer. doi:10.1007/978-1-4419-0146-0_8 .

Nunnally, J. C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Olivola, C. Y., & Oppenheimer, D. M. (2008). Randomness in retrospect: Exploring the interactions between memory and randomness cognition. *Psychonomic Bulletin & Review*, *15*(5), 991–996.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Ramsey, F., & Schafer, D. (2002). *The Statistical Sleuth: A Course in Methods of Data Analysis* (2nd ed.). Pacific Grove, CA: Duxbury.

Revelle, W. (2017). Psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illnois, USA, http://CRAN.R-project.org/package=psych Version=1.7.5

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154.

Rittle-Johnson, B., & Alibali, M. W. (1999). Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology*, *91*(1), 175.

Rossman, A., Chance, B., & Lock, R. (2001). *Workshop Statistics*. New York: Key College Publishing.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688.

Rubin, A., Bruce, B., & Tenney, Y. (1990). Learning about sampling: Trouble at the core of statistics. In *Proceedings of the third international conference on teaching statistics*. Dunedin, New Zealand.

Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Computers & Education*, *58*(1), 136–153. doi:10.1016/j.compedu.2011.07.017

Sabbag, A. (2013). *A psychometric analysis of the Goals and Outcomes Associated with Learning Statistics (GOALS) instrument* (Unpublished master's thesis). University of Minnesota, Minneapolis, MN.

Sabbag, A. & Zieffler, A. (2015). Assessing learning outcomes: An analysis of the GOALS-2 instrument. *Statistics Education Research Journal, 14* (2).

Saldanha, L. A. & Thompson, P. W. (2002). Conceptions of sample and their relationship to statistical inference. *Educational Studies in Mathematics, 51*, 257-270.

Santrock, J. W. (2011). *Educational psychology* (5th ed.). New York: McGraw Hill.

Sawilowsky, S. S. (2004). Teaching random assignment: Do you believe it works? *Journal of Modern Applied Statistical Methods, 3*(1), 221-226.

Singer, E., Hoewyk, J. V., & Maher, M. P. (2000). Experiments with incentives in telephone surveys. *Public Opinion Quarterly, 64*, pp. 171-188.

Smetana, L. K., & Bell, R. L. (2012). Computer simulations to support science instruction and learning: A critical review of the literature. *International Journal of Science Education*, *34*(9), 1337–1370. doi:10.1080/09500693.2011.605182

Smith, T. M., & Hjalmarson, M. A. (2013). Eliciting and developing teachers' conceptions of random processes in a probability and statistics course. *Mathematical Thinking and Learning*, *15*(1), 58–82. doi:10.1080/10986065.2013.738378

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, *36*, 404–411.

Tennyson, R. D., & Cocchiarella, M. J. (1986). An empirically based instructional design theory for teaching concepts. *Review of Educational Research*, *56*(1), 40–71. doi:10.3102/00346543056001040

Thompson, S. (2002). *Sampling* (2nd ed.). New York: John Wiley & Sons.

Thorndike, R. M., & Thorndike-Christ, T. (2010). Measurement and evaluation in psychology and education (8th ed.). Boston, MA: Pearson.

Tintle, N., Topliff, K., VanderStoep, J., Holmes, V. L., & Swanson, T. (2012). Retention of statistical concepts in a preliminary randomization-based introductory statistics curriculum. *Statistics Education Research Journal*, *11*(1), 21–40.

Triola, M. F. (2006). *Elementary Statistics* (10th ed.). Boston: Pearson Addison Wesley.

Utts, J. (2003). What educated citizens should know about statistics and probability. *The American Statistician*, *57*(2), 74–79.

Utts, J., & Heckard, R. (2007). *Mind on statistics* (3rd ed.). Cengage Learning.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction*, *4*(1), 45–69.

Vosniadou, S. (2012). Reframing the classical approach to conceptual change: Preconceptions, misconceptions and synthetic models. In B.J. Fraser, K. Tobin, & C.J. McRobbie (Eds.), *Second international handbook of science education* (pp. 119-128).

Vosniadou, S. (2013). Conceptual change in learning and instruction: The framework theory approach. In Vosniadou, S. (Ed.), *International handbook of research on conceptual change* (2nd ed., pp. 11-30). New York: Taylor & Francis.

Vosniadou, S., & Brewer, W. F. (1987). Theories of knowledge restructuring in development. *Review of Educational Research*, *57*(1), 51–67. doi:10.2307/1170356

Vosniadou, S., & Brewer, W. F. (1994). Mental models of the day/night cycle. *Cognitive Science*, *18*(1), 123–183.

Wagenaar, W. A. (1991). Randomness and randomizers: Maybe the problem is not so big. *Journal of Behavioral Decision Making, 4*(3), 220-222.

Wagler, A. & Wagler, R. Randomizing roaches: Exploring the "bugs" of randomization in experimental design. *Teaching Statistics, 36*(1), 13-20.

Wansink, B., Van Ittersum, K., & Painter, J. E. (2006). Ice cream illusions: bowls, spoons, and self-served portion sizes. *American journal of preventive medicine*, *31*(3), 240-243.

White, R. F. (1975). Randomization and the analysis of variance. *Biometrics*, *31*(2), 555–571. doi:10.2307/2529437

Wroughton, J.R., McGowan, H.M., Weiss, L.V., & Cope, T.M. (2013). Exploring the role of context in students' understanding of sampling. *Statistics Education Research Journal, 12*(2), 32-58.

Wu, C. F. J., & Hamada, M. (2000). *Experiments: Planning, analysis, and parameter design optimization*. New York: Wiley.

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing groups: Randomization and bootstrap methods using R*. Hoboken, NJ: John Wiley & Sons.

Zieffler, A., & Catalysts for Change. (2013*). Statistical Thinking: A simulation approach to uncertainty* (2nd ed.). Minneapolis, MN: Catalyst Press.

Zieffler, A., & Catalysts for Change. (2015*). Statistical Thinking: A simulation approach to uncertainty* (3rd ed.). Minneapolis, MN: Catalyst Press.

Appendix A: Correspondence with students in EPSY 3264 course

Appendix A1: Invitation e-mail sent to students in the online section (EPSY 3264-004)

Subject: EPSY 3264: Invitation to Participate in Dissertation Research

Hello,

You are receiving this email because you are enrolled in EPSY 3264: Basic and Applied Statistics. I am inviting you to participate in a study I am conducting as part of my dissertation research in Statistics Education in the Quantitative Methods in Education program in the Department of Educational Psychology at the University of Minnesota.

The study has the objective of finding ways to improve introductory statistics students' understanding of concepts related to study design and conclusions. While statistics education guidelines indicate that these concepts are essential for students to understand, prior research in statistics education shows that students often struggle with these concepts even after completion of an introductory statistics course. However, little research exists on effective ways to improve understanding of study design and conclusions. This research will add needed information about best practices in teaching about the connections between study design and inferences that can be made. Your help with this is greatly appreciated. By agreeing to participate, you would give me permission to use your IDEA, Quiz #5, and Lab Assignment #8 answers that are part of your work in the EPSY 3264 Basic and Applied Statistics course for research purposes only. All information that could identify you will be removed for the analysis.

If you decide to participate in the study, no specific action is needed from you at this point. If you decide not to participate, please reply to this email saying no and you will not be included in the study. Not participating in this study will not affect your grade in EPSY-3264 in any way. Participation is voluntary.

Attached you will find the consent form for this study, please review it before you decide about your participation in this study. (You may ignore the sentence about the class being videotaped, as that does not apply online.)

If you have any questions or concerns regarding this study, please contact me at fryxx069@umn.edu.

Sincerely

Elizabeth Fry

Doctoral candidate in Educational Psychology (QME)
Department of Educational Psychology
Room 161 Educational Sciences building
56 East River Road
Minneapolis, MN 55455

Appendix A2: Consent form given to all EPSY 3264 students

INFORMATION SHEET FOR RESEARCH

Students' Conceptual Understanding of Study Design and Conclusions

You are invited to be in a research study where the aim of this study is to explore the impact of a five-day introductory statistics class unit about study design and conclusions on students' understanding of these concepts. You were selected as a possible participant because you are enrolled in the course EPSY 3264: Basic and Applied Statistics. We ask that you read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by Elizabeth Fry, PhD Candidate in the Department of Educational Psychology at the University of Minnesota.

Procedures:

If you agree to be in this study, I would ask for permission to use your answers on Group Quiz #5, Lab Assignment #7 Part 1, and Lab Assignment #8 answers that are part of your work in the EPSY-3264 Basic and Applied Statistics course for research purposes only. All information that could identify you will be removed for the analysis.

The four class activities and group quiz that are part of this unit will be observed by the researcher (myself, Elizabeth Fry) and a fellow graduate student co-observer. I will also videotape the class, with the focus being on the instructor. The recordings and observations will be for research purposes only, and only I will view the recordings. The recordings will be destroyed after data analysis.

Confidentiality:

The records of this study will be kept private. In any sort of report that might be published, no information will be included that will make it possible to identify a subject. Research records will be stored securely and only the researcher will have access to the records.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University of Minnesota, nor will it affect your grade in EPSY 3264. If you decide not to participate, you are free to withdraw at any time without affecting those relationships.

Contacts and Questions:

The researcher conducting this study is Elizabeth Fry. You may ask any questions you have now. If you have questions later, you are encouraged to contact her at Room 161 EdSciB, 56 E River Road, Minneapolis, MN 55455, phone: 612-624-1099, fryxx069@umn.edu. You may also contact her academic advisor Professor Robert C. delMas, phone: 612-625-2076, email: delma001@umn.edu

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher(s), you are encouraged to contact the Research Subjects' Advocate Line, D528 Mayo, 420 Delaware St. Southeast, Minneapolis, Minnesota 55455; (612) 625-1650.

You will be given a copy of this information to keep for your records.

Appendix B: Activities and readings: in-class versions

Appendix B1: Sampling Countries activity

Course Activity: Sampling Countries

In this activity, you will compare different ways of taking samples of countries of the world from a population of countries.

1. Think of 20 countries that you believe are representative of the countries in the world (i.e., they resemble the collection of all countries of the world). Fill in the list of countries in the table below.

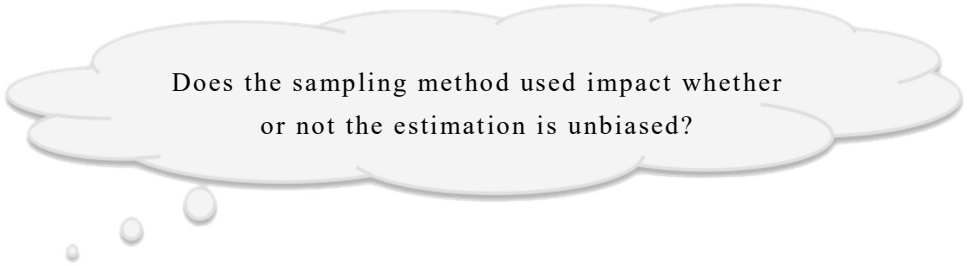
Country
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.

2. Describe how your list of countries is representative of the countries of the world.

In this activity, you will have access to a population of 196 countries of the world and some information about their life expectancy as determined by the World Bank (www.worldbank.org) in 2013. The data can be found in the last few pages of this activity. (Please note that not quite all of the countries of the world are in this dataset because some had missing data, but we will consider this list of 196 countries to be our *population* of countries.) You will examine the following variable of interest:

Life Expectancy: The number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

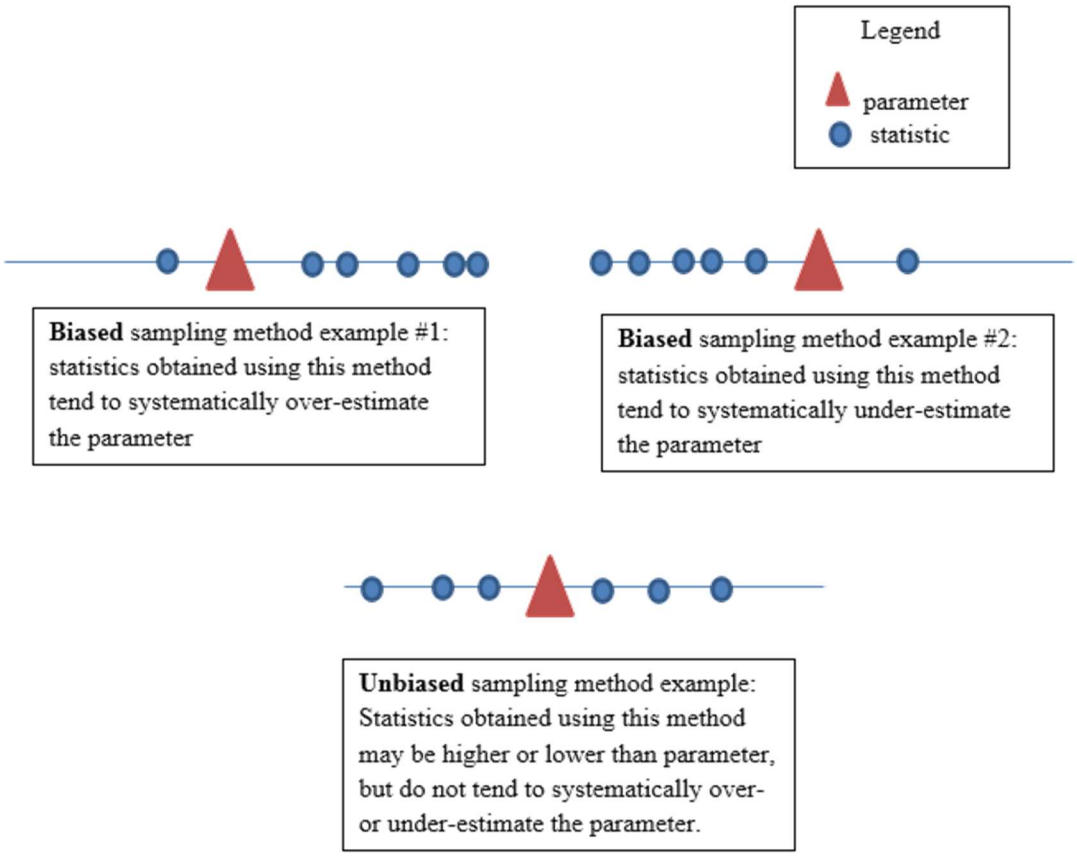
In this activity, you will be exploring the following research question:



Unbiased Estimation

One concern when taking a sample is whether or not an estimate taken from a sample (**statistic**) will appropriately estimate the "truth" of the population (**parameter**). When a sampling method produces statistics that tend to systematically over- or under-estimate the population parameter, we call that sampling method **biased.** Ideally, we want sample estimates to be **unbiased**. Unbiasedness means that the estimation method used tends to produce sample statistics that are around the population parameter, without the tendency to over-estimate or under-estimate the parameter.

For example, as illustrated in the picture below, suppose we are trying to estimate a parameter of the population, symbolized by a triangle. Statistics taken from different samples will vary, as symbolized by the small circles. The biased sampling method examples show how biased methods produce estimates that that tend to be higher or lower than the parameter we are trying to estimate. In contrast, the unbiased sampling method example shows how some estimates are on the low side, some estimates are on



the high side, but as a whole they are centered on the true value of the parameter.

Follow these instructions to compute and report the average life expectancy for your sample of countries:

- Open up a blank TinkerPlots™ file.
- Drag a Table from the Object toolbar into your document.
- Create a new attribute called *Life Expectancy* in the first column of the case table.
- Using the tables at the back of this activity for reference, enter the life expectancies of your 20 countries under the *Life Expectancy Column*.
- Plot the 20 life expectancies.
- Separate and stack the cases, then find the value of the Average.

3. Write down the value of the average life expectancy of your 20 countries here.

4. Is this value a parameter, or a statistic?

Add this sample estimate to the case table on the instructor's computer.

5. Sketch the plot of all of the sample average life expectancies. Make sure to label the axes appropriately.

The population average life expectancy of all 196 countries (parameter) turns out to be **71 years.** On your plot above, draw a vertical reference line marking this value.

6. Were most of your sample estimates around the population average?

7. Approximately what percentage of groups in your class had sample statistics that exceeded the population average?

8. Based on your answers to the previous two questions, does this method of sampling appear unbiased, or does it tend to over-estimate or under-estimate the average life expectancy of the population of countries?

9. What are some reasons for why the sampling method of asking people to name 20 countries might produce biased estimates?

In order to try to eliminate potential biases that can occur by human selection, it is better to take a **random sample**. Humans are not very good "random samplers" – even though we are trying to obtain a representative sample, we tend to name countries that are more well known or appear more often in the news than others. Instead, it is important to use random sampling techniques to do the sampling for us.
The goal of random sampling is to obtain a representative sample, so estimates of population parameters are unbiased. Although there is variation from sample to sample, there is no systematic tendency to over-estimate, or to under-estimate, the population parameter.

Simple Random Sampling

A simple random sample (SRS) is a specific type of random sample that gives every observational unit in the population the same chance of being selected. In fact, every sample of size *n* has the same chance of being selected. In this example, we will take a simple random sample of 10 countries.
The first step in drawing a simple random sample is to obtain a **sampling frame**, which is a list of each member of the population (in this case, this will be a list of all of the countries in our population). We have already prepared a sampling frame of the countries for you and saved it in the *SamplingCountries.tp* file.

USE TINKERPLOTS™ TO DRAW A SIMPLE RANDOM SAMPLE

- Open the file SamplingCountries.tp
- Draw one simple random sample of 10 countries from the sampler. (Note that the sampler has been set up to draw the sample without replacement so you do not get any duplicates.)

First, you will examine the distribution of life expectancy for this single sample.

10. Plot the "Life Expectancy" variable for this single sample. Sketch a plot below, being sure to label the axes.

11. Obtain and record the average life expectancy for this single sample.

12. Compare this plot and average to the plot and average obtained by another group near you. Did you get the exact same plot and sample average? Are they similar?

13. Now, compare your average from this sample to the population average of 71 years. Are the averages the same? Are they similar?

You may have noticed that your sample differed from other samples taken by your classmates. Samples differ, but hopefully, your sample estimate should be somewhat close to the population average life expectancy, if it is a representative sample.

Now, we will investigate whether random sampling produces sample estimates that are unbiased.

14. If we took many *random* samples of size 10 and made plots of the sample average life expectancies similar to the plot you drew from your instructor's computer in question #5, what do you think this plot would look like?

15. Where do you predict this plot will be centered?

In TinkerPlots™, go back to the plot of the life expectancies from the sample of size 10 you just examined.

- Collect the average life expectancy from your random sample.
- Carry out 200 trials of the simulation.
- Plot the 200 average life expectancies you collected.
- Obtain the average from your plot of the 200 sample statistics.

16. Sketch the plot of these 200 averages and make sure to label the axes appropriately.

17. If the sampling method is unbiased, where should you expect the plot to be centered?

18. Is your plot centered near that value?

19. Based on your answer to the previous question, does simple random sampling produce an unbiased estimate of the average country's life expectancy? Explain.

20. Compare your plot above in question #16 with the plot you made in question #5. What do you think is better: taking a larger convenience sample (n = 20), or taking a smaller, random sample (n = 10)? Explain your choice.

21. When you draw a single random sample from a population, do you expect your sample statistic to match the population parameter *exactly?* Why or why not?

22. What does it mean for a sampling method to be *unbiased?*

Because random sampling is an unbiased sampling method, it allows us to use our samples to make generalizations, or wider inferences, about the population from which the sample was taken.
In real studies, researchers do not have access to information about the full population like you did in this activity. However, they need to use a sampling method that tends to produce representative samples that give unbiased estimates, so

that they can make valid generalizations to the population of interest. For example, if a researcher took a random sample of countries from this population and found the sample average life expectancy to be 72.5, (s)he could generalize that the average country's life expectancy from this population is approximately around 72.5.

In statistics, **estimation** refers to the process by which one makes inferences about a population or model, based on information obtained from a sample. The **population** is the entire collection of who or what (e.g., the observational units) that you would like to draw inferences about. In practice, it is often impossible to examine every unit of the population, so data from a subset, or **sample**, of the population is examined instead. The sample data provides statisticians with the best estimate of the exact "truth" about the population. The "truth" one is searching for in the population is typically a summary measure such as the population average or population percentage. Summary measures of a population are called **parameters**. The estimates of these values from sample data are referred to as **statistics**

Country Name	Life Expectancy
Afghanistan	60.03
Albania	77.54
Algeria	74.57
Angola	51.87
Antigua and Barbuda	75.78
Argentina	75.99
Armenia	74.56
Aruba	75.33
Australia	82.20
Austria	80.89
Azerbaijan	70.69
Bahamas, The	75.07
Bahrain	76.55
Bangladesh	71.25
Barbados	75.33
Belarus	72.47
Belgium	80.39
Belize	69.98
Benin	59.31
Bermuda	80.57
Bhutan	69.10
Bolivia	67.91
Bosnia and Herzegovina	76.28
Botswana	64.36
Brazil	74.12
Brunei Darussalam	78.55
Bulgaria	74.47
Burkina Faso	58.24
Burundi	56.25
Cabo Verde	72.97
Cambodia	67.77
Cameroon	55.04
Canada	81.40
Central African Republic	49.88
Chad	51.19
Channel Islands	80.46
Chile	81.20
China	75.35
Colombia	73.81

Country Name	Life Expectancy
Comoros	62.93
Congo, Dem. Rep.	58.27
Congo, Rep.	61.67
Costa Rica	79.23
Cote d'Ivoire	51.21
Croatia	77.13
Cuba	79.26
Cyprus	79.95
Czech Republic	78.28
Denmark	80.30
Djibouti	61.69
Dominican Republic	73.32
Ecuador	75.65
Egypt, Arab Rep.	70.93
El Salvador	72.50
Equatorial Guinea	57.29
Eritrea	63.18
Estonia	76.42
Ethiopia	63.44
Faeroe Islands	81.30
Fiji	69.92
Finland	80.83
France	81.97
French Polynesia	76.33
Gabon	63.84
Gambia, The	60.00
Georgia	74.08
Germany	81.04
Ghana	61.14
Greece	80.63
Grenada	73.19
Guam	78.87
Guatemala	71.49
Guinea	58.22
Guinea-Bissau	54.84
Guyana	66.31
Haiti	62.40
Honduras	72.94
Hong Kong	83.83
Hungary	75.27

Country Name	Life Expectancy
Iceland	83.12
India	67.66
Indonesia	68.70
Iran, Islamic Rep.	75.13
Iraq	69.47
Ireland	81.04
Israel	82.06
Italy	82.29
Jamaica	73.47
Japan	83.33
Jordan	73.90
Kazakhstan	70.45
Kenya	60.95
Kiribati	65.77
Korea, North	69.79
Korea, South	81.46
Kuwait	74.46
Kyrgyz Republic	70.20
Lao PDR	65.69
Latvia	73.98
Lebanon	80.13
Lesotho	49.33
Liberia	60.52
Libya	71.66
Liechtenstein	82.38
Lithuania	74.16
Luxembourg	81.80
Macao SAR, China	80.34
Macedonia, FYR	75.19
Madagascar	64.67
Malawi	61.47
Malaysia	74.57
Maldives	76.60
Mali	57.54
Malta	80.75
Mauritania	62.80
Mauritius	74.46
Mexico	76.53
Micronesia, Fed. Sts.	68.97
Moldova	68.81

Country Name	Life Expectancy
Mongolia	69.06
Montenegro	74.76
Morocco	73.71
Mozambique	54.64
Myanmar	65.65
Namibia	64.34
Nepal	69.22
Netherlands	81.10
New Caledonia	77.12
New Zealand	81.41
Nicaragua	74.51
Niger	60.83
Nigeria	52.44
Norway	81.45
Oman	76.84
Pakistan	65.96
Panama	77.42
Papua New Guinea	62.45
Paraguay	72.80
Peru	74.28
Philippines	68.13
Poland	76.85
Portugal	80.37
Puerto Rico	78.71
Qatar	78.42
Romania	74.46
Russian Federation	71.07
Rwanda	63.39
Samoa	73.25
Sao Tome and Principe	66.26
Saudi Arabia	74.18
Senegal	65.88
Serbia	75.14
Seychelles	74.23
Sierra Leone	50.36
Singapore	82.35
Slovak Republic	76.26
Slovenia	80.28
Solomon Islands	67.72

Country Name	Life Expectancy
Somalia	55.02
South Africa	56.74
South Sudan	55.22
Spain	82.43
Sri Lanka	74.24
St. Lucia	74.91
St. Vincent and the Grenadines	72.81
Sudan	63.17
Suriname	70.99
Swaziland	48.94
Sweden	81.70
Switzerland	82.75
Syrian Arab Republic	74.72
Tajikistan	69.40
Tanzania	64.29
Thailand	74.25
Timor-Leste	67.52
Togo	59.13
Tonga	72.64

Country Name	Life Expectancy
Trinidad and Tobago	70.31
Tunisia	73.65
Turkey	75.18
Turkmenistan	65.46
Uganda	57.77
Ukraine	71.16
United Arab Emirates	77.20
United Kingdom	80.96
United States	78.84
Uruguay	76.84
Uzbekistan	68.23
Vanuatu	71.67
Venezuela, RB	74.07
Vietnam	75.76
Virgin Islands (U.S.)	79.62
West Bank and Gaza	73.20
Yemen, Rep.	63.58
Zambia	59.24
Zimbabwe	55.63

Appendix B2: Establishing Causation reading

Establishing Causation

Researchers often examine relationships between variables. Two variables are associated if the values of one variable tend to be related to the values of another variable. In particular, an **explanatory variable** is a variable that can be used to help us understand or predict values of the **response variable**.

In many studies, the goal is more than to determine an association. The goal is to determine whether changes in an explanatory variable influence, or cause, changes in a response variable. However, association does not necessarily mean that there is a **cause-and-effect** relationship: namely, that changing the values of one variable will influence the value of another variable. Consider this example:

Suppose educators are trying to figure out if taking a test preparation class will increase students' test scores. Students are allowed to choose whether to take the class or not, and in the end, the data show that the students who took the class scored significantly higher on the test than the students who did not ($p < .05$).

Here, the explanatory variable is whether or not the students took the class, and the response being measured is the test score. The researchers found a significant association between these variables.

But can the researchers conclude that the test preparation class was effective? Not necessarily. Think about how students who chose to take the class might be different from students who chose not to take it. Perhaps the students who chose to take the class would have had higher scores even if they had not taken the class, just because they're already more motivated to succeed or have higher GPAs than students who did not take the class. In this case, students' motivation and GPA are called **confounding variables** because they help to offer a plausible explanation for the observed association.

A study where researchers do not manipulate the explanatory variable is called an **observational study.** In this type of study, researchers may compare groups, but do not control which group a participant is in. The exam preparation class scenario above is a good example of an observational study,

because the subjects choose whether or not to take the class. The researcher did not control this. The problem with observational studies is that cause-and-effect conclusions are difficult to make because the groups of participants being compared may differ in ways other than the explanatory variable, and confounding can come into play.

In contrast, in an **experiment**, the researcher actively has control over which group each subject is in. When categories of the explanatory variable are assigned to subjects in an experiment, the explanatory variable is also called a **treatment variable. (**Recall that you have already seen examples of treatment variables in course activities such as *Memorization* and *Sleep Deprivation*.)

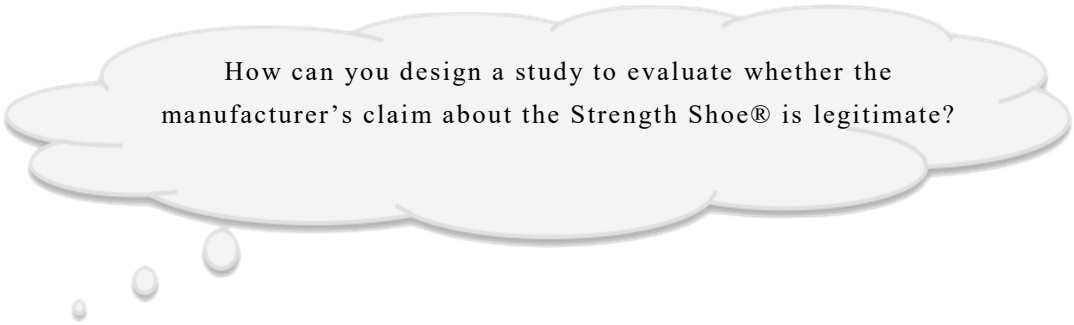
Consider the above example of the test preparation class. If you were to assign students to take the class or not, how would you do this? It's important to try to make sure that students who are more motivated, have higher GPAs, or study longer, are *not* more likely to end up in one group than the other. If the students in the class were similar in all respects to the students who did not take the class, then if we found that students who took the class did significantly better on the test, we could argue this was because of the class. Since the only major difference between the groups is that one took the class and one did not, we can argue that the class led to the higher scores.

As we will see in the next activity, **random assignment** is a method to create groups that are similar in all respects except for the treatment imposed. Random assignment will not produce groups that are *exactly* equivalent to each other with respect to *every* possible confounding variable. However, assigning randomly means that subjects with certain characteristics will *not* be more likely to be in one group than the other. The goal is to create similar groups, so we can argue that any observed significant differences in the response variable are because of the only major difference between the groups: the treatment variable. Therefore, using random assignment has the potential to allow researchers to establish a *cause-and-effect* relationship between the explanatory and response variables.

Appendix B3: Strength Shoe activity

Course Activity: Strength Shoe®

The Strength Shoe® is a modified athletic shoe with a 4-cm platform attached to the front half of the sole. Its manufacturer claims that people who wear this shoe can jump farther than people who wear ordinary training shoes. In this activity you will be



examining the following question:

Discuss the following questions.

Suppose that you take a random sample of individuals by randomly selecting them from the population. You observe who does and does not wear the Strength Shoe®, and then compare the two groups' jumping ability.

1. Why would it be advantageous to take a random sample of individuals for this study?

Suppose you find in this study that on average, the group who wears strength shoes can jump much farther than the group who wears ordinary training shoes.

2. Do you think this is compelling evidence that strength shoes really increase jumping ability? Explain.

An association may not necessarily point to a cause-and-effect relationship. For example, subjects who choose to wear the Strength Shoe® could be more athletic to begin with than those who opt to wear the ordinary training shoes, and this is why they can jump farther.

When researchers want to find if a treatment variable *causes* changes in a response, they control the treatment variable and assign study participants to treatment groups. What we want to see is that both groups are approximately equal in terms of other variables, such as athletic ability, height, sex, etc. so that these other variables are *not* causing differences in jumping ability. We want to be able to say that the type of shoe is what is affecting jumping ability.

A 1993 study published in the *American Journal of Sports Medicine* investigated the Strength Shoe® claim using 12 intercollegiate track and field athletes as study participants[1]. Suppose you also want to investigate this claim, and you recruit 12 of your friends to serve as subjects. You plan to have six people wear a Strength Shoe® and the other six wear the ordinary training shoes they usually wear when exercising, and then measure each group's jumping ability.

Confounding Variables

Two factors that might affect jumping distance are a person's sex and height. In every study, there are potentially many factors (aside from the treatment) that may be related to the response variable and, in turn, affect the results of the study. Statisticians refer to these variables as **confounding variables**.

One potential way to deal with this issue would be to purposefully try to balance out certain confounding variables and create two groups that are relatively equivalent with respect to known confounding variables. Suppose the researcher decided to control for sex and height by purposefully assigning the two groups so that there was an equivalent number of females in each group, and the average height for each group was roughly equivalent:

---

[1] Cook, S. D., Schultz, G., Omey, M. L., Wolf, M. W., & Brunet, M. F. (1993). Development of lower leg strength and flexibility with the strength shoe. *American Journal of Sports Medicine, 21*, 445–448.

Ordinary Training Shoe Group		
Name	Sex	Height
Jasmine	Female	61
Ka Nong	Female	67
George	Male	67
Paul	Male	73
Tong	Male	71
Ringo	Male	71

Strength Shoe Group		
Name	Sex	Height
Keyaina	Female	63
Mary	Female	66
Antonio	Male	68
Andreas	Male	70
Davieon	Male	70
John	Male	69

3. Based on the tables above, does it appear that the two groups are equivalent to each other with respect to the Sex variable? Explain.

Now, we will compare the two groups in the above sample based on height.
Open the file *StrengthShoe-Purposeful.tp*
Next, plot the height variable as follows:

- Plot the attributes Height (*x*-axis) and Group (*y*-axis) in a single plot.
- Separate and stack the cases.
- Display the average for each group.

4. Examine your plot of heights and write down the average height for each of the two groups. Also, compute the difference in the two averages. Are the two groups roughly equivalent with respect to height?

Average Height of Strength Shoe Group: _____

Average Height of Ordinary Training Shoe Group: _____

Difference in average height (Strength Shoe®– Ordinary Training Shoe):

If the two groups are balanced with respect to sex and height, then if you find a significant difference in jumping ability between the two groups, you can argue that it was not differences in sex or height that caused the difference in jumping ability.

5. Can you think of other variables besides sex and height that might explain differences in jumping ability? If so, what are they?

Suppose now that there is a genetic factor (which you did not measure before the study) that will strongly influence how far participants will be able to jump (regardless of the shoe type). Let's call it the "*X*-factor." Since you do not know about it, you have no way to measure and control for it, but it will likely influence the results of our study. For example, what if more participants assigned to the Strength Shoe® group have this *X*-factor? Then the Strength Shoe® group would show increased jumping ability, even if training with a StrengthShoe® is no better than training with an ordinary training shoe.

- To explore this, we actually have an *X*-Factor already hidden in the TinkerPlots™ file! To show it, right click anywhere in the table you see in the TinkerPlots™ window and select Show Hidden Attribute.

You will now see that the *X*-Factor variable ("Yes" or "No", indicating whether the participant has that genetic factor or not) has appeared as another attribute in the trial results. It is important to remember that in real life, you would not know about this confounding variable. But, here, you can examine how this confounding variable is distributed between the Strength Shoe® and Ordinary Training Shoe groups when the conditions are purposefully assigned.

- Plot the attributes X-Factor (*x*-axis) and Group (*y*-axis) in a single plot.
- Display the percentages for each group.

6. Calculate the percent of the subjects in each group that have the *X*-factor. Do you think the two groups are roughly equivalent with respect to whether or not they have the *X*-factor? Explain.

Percent of people with *X*-factor in Strength Shoe Group: _____

Percent of people with *X*-factor in Ordinary Training Shoe Group: _____

Difference in percent with *X*-factor (Strength Shoe®– Ordinary Training Shoe):

Suppose the 12 subjects were purposefully assigned to control for sex and height, as above. Researchers find that the subjects wearing the Strength Shoes® jump significantly farther, on average, than the subjects wearing ordinary training shoes.

7. Do you think we could conclude that the shoes caused the difference in jumping ability?

While some confounding variables may be identified and controlled in a study, others may not be identified initially by the researcher, such as the *X*-factor in this example. It is impossible to know about and observe all possible confounding variables.

Luckily, it turns out that the key to controlling for *all* of these confounding variables (both observed and unobserved) is to use *random assignment* in forming experimental groups.

Random Assignment

Random assignment is the preferred method of assigning subjects to treatment conditions in an experiment. Under random assignment, each subject has an equal chance (probability) of being assigned to any of the treatment conditions.

Observed Variable: Height

Next you will use TinkerPlots™ to randomly assign subjects to the two groups. Then, we want you to examine the average height in each group to see if height differences in the two groups could explain the jumping differences we saw between the groups.

- Open the *Strength-Shoe-Random-1.tp* TinkerPlots™ file.

Note that the model has already been set up for you; there is a Counter device with the study participants and a Stacks device that will randomly assign each participant to a group.

- Press Run to record the results of a single random assignment.
- Plot the attributes Height (x-axis) and Group (y-axis) in a single plot;
- Separate and stack the cases.
- Display the average for each group.

8. Calculate the average height for each group. Also find the difference in these two averages (taking the Strength Shoe® group's average minus the ordinary training shoe group's average).

   Average height in Strength Shoe® Group: _____

   Average height in Ordinary Training Shoe Group: _____

Difference in average height (Strength Shoe®– Ordinary Training Shoe):

9. In this single random assignment, are the two groups exactly balanced with respect to height? Explain.

This is just a single random assignment, and we want to get a sense of the difference in the average height across many random assignments.

10. Suppose you want to make a plot of the difference in average height for many different random assignments. Where do you predict this plot will be centered?

Now, construct this plot as follows:

- Use the Ruler tool to compute the difference in the average heights between the two groups. (Note: Subtract the Ordinary Training Shoe group from the Strength Shoe® group.)
- Right click on the difference in averages and select Collect Statistic.
- Collect 499 more trials.
- Plot the 500 differences.
- Organize and fully separate the results (no bin lines) for the plot.

- Show the Average (and its numeric value) on the plot.

11. Sketch the plot below.

12. Where is this plot centered?

13. What does your answer to the previous question imply about the tendency of random assignment to balance out the height variable in the two groups? Explain.

Unobserved Confounding Variable

As we saw earlier, the variable *X*-Factor is an unobserved confounding variable (or lurking variable) that the researcher does not observe, but will strongly influence how far participants can jump, regardless of the shoe they use.

- Open the *Strength-Shoe-Random-2.tp* TinkerPlots™ file found on Moodle and Run the Sampler.
- Again, we actually have an *X*-Factor already hidden in the TinkerPlots™ file! To show it, right click anywhere in the table of trial results and select Show Hidden Attribute.

- Plot the attributes X-Factor (*x*-axis) and Group (*y*-axis) in a single plot.
- Display the percentage for each group.

14. Calculate and report the percent of people with the *X*-Factor in each group. Also subtract these two percentages (subtract the ordinary training shoe group's percent from the Strength Shoe® group's percent).

    Percent of people with the *X*-Factor in Strength Shoe® Group: _____

    Percent of people with the *X*-Factor in Ordinary Training Shoe Group: _____

Difference in percentages of people with *X*-Factor (Strength Shoe®– Ordinary Training Shoe):

Again, this is just a single random assignment and we want to get a better sense of the differences in the *X*-Factor across many random assignments to groups.

15. Suppose you want to make a plot of the difference in percentages of people with the *X*-Factor across many random assignments. Where do you predict this plot will be centered?

We need to again collect *two measures*: the percentage of participants with the *X*-Factor in the Strength Shoe® group and the percentage of participants with the *X*-Factor in the Ordinary Training Shoe group.

Create a plot of the differences in percentage of people with the *X*-Factor for each group as follows:

- Right-click on the percent with the *X*-Factor for the Strength Shoe® group and select Collect Statistic.
- Right-click on the percent with the *X*-Factor for the Ordinary Training Shoe group and select Collect Statistic.
- Create a third attribute in your History of Results table and name it *Difference*.
- Right-click *Difference* and select Edit Formula.
- Use the Formula Editor to compute the difference in the percent of people with the *X*-Factor between the two groups. (Note: Subtract the Ordinary Training Shoe group from the Strength Shoe® group.)
- Collect 499 more trials.
- Plot the 500 differences.
- Organize and fully separate the results (no bin lines) for the plot.
- Show the Average (and its numeric value) on the plot.

16. Sketch the plot below.

17. Where is this plot centered?

18. What does your answer to the previous question imply about the tendency of random assignment to balance out the *X*-factor variable in the two groups? Explain.

Conclusions

19. Why was it important to look at the *X*-factor in this study, rather than just focusing on sex and height?

20. Which method of assignment is better: Purposefully assigning groups to balance out known confounding variables, or assigning groups randomly? Explain.

Suppose we conduct this random assignment and find that the Strength Shoe® group jumps significantly farther, on average, than the ordinary training shoe group.

21. Would you be comfortable concluding that Strength Shoes® caused the increased jumping distance? How would you argue that most likely no confounding variable was responsible for this difference in jumping distance?

22. Now, look back at how the actual sample of 12 subjects was collected back on the third paragraph of page 2. Would you be comfortable generalizing the results of a study based on that sample to conclude that training with Strength Shoe® will increase jumping distance for all athletes? Explain.

Appendix B4: Scope of Inferences Reading

Scope of Inferences

The inferences one can draw from a statistical study depend on how the study was designed. There are two types of inferences researchers may wish to draw from a study: (1) generalization to a population and (2) making cause-and-effect conclusions. These are two distinct types of conclusions, and randomness plays a different role in the study design for each.

Generalization to a Population

In statistical studies, we often wish to draw a conclusion about a population of interest, using a sample drawn from that population. In other words, we wish to **generalize** our results back to our population of interest. To generalize means to make a claim about a wider population of interest, using a sample of data. In order to do this, we need a representative sample, or one that is similar in characteristics to the population. Consider this example:

The Physician's Health Study[2] *was conducted in the 1980's to study whether or not taking a daily low-dose aspirin reduced the risk of heart attacks. The sample was gathered by initially sending out letters to recruit male physicians between the ages of 40 and 84 who lived in the United States and who were registered with the American Medical Association. Using a sample of over 30,000 willing and eligible physicians, an experiment was conducted and it was found that the subjects who took the low-dose aspirin were significantly less likely to suffer from heart attacks than those who took a placebo.*

Can we say that for all adults in the U.S., those who take aspirin daily are less likely to suffer from heart attacks than those who do not? This would be making a **generalization** to the population of U.S. adults.

Given how the sample was taken, there could be **bias** in estimating the difference in heart attack rates between adults who take daily aspirin and adults who do not. This is because the sample of U.S. male physicians is likely not representative of all U.S. adults. Females are not represented. It's quite possible that male physicians have diets, exercise routines, and other health habits that are different from those of the general adult population. It is difficult to reliably make any generalizations about aspirin and heart attacks to a wider population of U.S. adults when the sample is a biased representation of this population.

In order to avoid this bias, the best way to obtain a representative sample is **random sampling**. By using randomness to select subjects, we eliminate human bias that makes some units more likely to be in the sample than others. Instead, we give all units in the population an equal chance to be in the sample. By sampling randomly, we are likely to get a sample that looks more or less like a snapshot of the population. An **unbiased** sampling method like random sampling means that we will not have a tendency to over-

[2] http://phs.bwh.harvard.edu/

estimate or under-estimate the parameter of interest. Then, we can use the sample to make generalizations about the population that it represents.

Establishing Causation

Making a cause-and-effect conclusion is a separate goal that may be desired from a study. In order to make causal claims, a significant association must first be found between a treatment variable and a response variable. If the two variables are associated, we can conclude there is a relationship, but we cannot necessarily conclude that changes in the treatment variable will lead to changes in the response variable. However, when we establish **causation**, we claim that changes in the treatment variable influence the value of the response variable.

In the example above with the Physician's Health Study, an association was observed: those who took daily aspirin were less likely to suffer from a heart attack than those who took a placebo. But can we conclude that the aspirin was the *cause* of the lower heart attack rates?

It's important to first consider whether any **confounding** variables – that is, variables that may be associated with both the treatment and response variables – could be responsible for the observed association. If the physicians were allowed to self-select whether they took aspirin or not, it's possible that those who chose to take aspirin might have tended to have healthier lifestyles than those who did not take any aspirin. Then, it could be the difference in tendency to live a healthy lifestyle – not the aspirin itself – which might have been responsible for the difference in heart attack rates.

However, in the Physician's Health Study on aspirin, the subjects were **randomly assigned** into two groups: one took aspirin and one took the placebo. With this random assignment, we are *not* more likely to get subjects with healthier lifestyles in one group than another – everyone has an equal chance to be in each group. The random assignment tends to balance out the groups with respect to all potential confounding variables. If the only major difference between the groups was that one took aspirin and the other one took a placebo, then any differences in heart attack rates observed can be attributed to the aspirin vs. placebo treatment. Therefore, the fact that those who took aspirin were less likely to develop a heart attack is evidence that the aspirin lowered the heart attack rates.

Two Types of Inferences

Note that *generalization to a population* and *establishing causation* are two different types of inferences. Randomness is a desired part of the study design for making each of these inferences, but the type of randomness we need for each inference should not be confused.

With generalization, we ask the question: "Can we use the results from this sample to make a broader claim about the population the sample was taken from?" To answer yes, we ideally need random sampling to enable a sample that is representative of the population. Random sampling is intended to reduce the differences between sample and population, so that we can generalize to this population.

With establishing causation, we ask the question: "Does changing one variable lead to a change in another variable?" To answer yes, we ideally need random assignment in order to balance out confounding variables between treatment groups so that they are similar in all respects except for the level of the treatment variable. Random assignment is

intended to reduce the differences between the two groups due to factors that are *not* being manipulated in the experiment, so that if there are differences in the response variable, we can attribute these differences to the treatment variable.

It is possible to have random sampling, random assignment, both, or neither. These different types of study designs and the inferences you can make from them are summarized in the table below.

		Selection of Units	
		Random Sampling	**No Random Sampling**
Allocation of Units to Groups	**Random Assignment**	Can make a causal conclusion and can generalize conclusion to the population.	Can make a causal conclusion but cannot generalize this conclusion to the population
	No Random Assignment	Can generalize to the population, but cannot make causal claims.	Cannot generalize to the population, and cannot make causal claims either.

In the case of the Physician's Health Study, we had random assignment, but not random sampling. We can claim that taking daily aspirin reduces the chance of heart attack, but this may only be true for men similar in characteristics to those in the sample. We cannot necessarily generalize this claim to all adults, or even all males ages 40-84. Males in the overall population may be different from these physicians in characteristics such as health and exercise habits, and thus may respond differently to aspirin than the physicians in this study.

Ideally, it would be great if we could have both random sampling and random assignment, so that we could make causal claims that we could generalize to the population. The reality in study design is that it is difficult and rare to have a study that uses both random sampling and random assignment. In order to perform an experiment with random assignment, often the study participants have to give up a lot of time. It is much easier to recruit people who are willing to give up their time and be in the study than to randomly sample from the population and rely on the people in this random sample to participate in the study.

In many cases, we have studies that have neither random sampling nor random assignment. With these studies, we cannot necessarily generalize findings to a population nor establish any causal claims. However, results from these studies may still reveal interesting findings and lead to further research.

Appendix B5: Murderous Nurse activity

Course Activity: Murderous Nurse

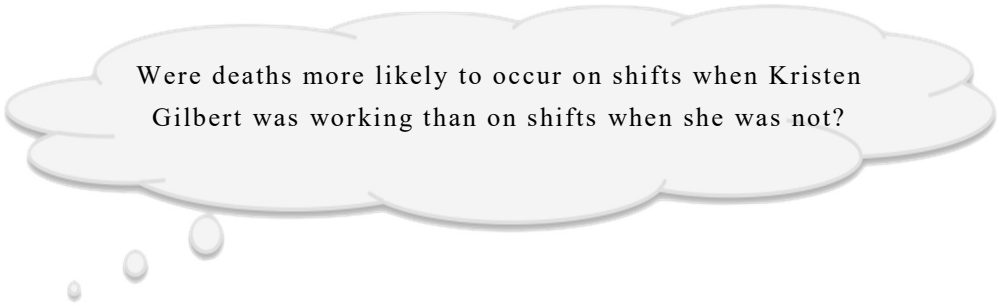
For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine.

Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU[3]. Here are the data presented during her trial:

	Gilbert working on shift	Gilbert not working on Shift	Total
Death occurred on Shift	40	34	74
No death occurred on shift	217	1350	1567
Total	257	1384	1641

You will use these data to answer the following research question

[3] Cobb, G. W., & Gehlbach, S. (2006). Statistics in the courtroom: United States vs. Kristen Gilbert. In R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern (Eds.), *Statistics: A guide to the unknown* (4th Edition), pp. 3–18. Duxbury: Belmont, CA.


Were deaths more likely to occur on shifts when Kristen Gilbert was working than on shifts when she was not?

Discuss the Following Questions

1. Among all 1,641 shifts, what percentage of shifts had a death occur?

2. Among the 257 shifts when Gilbert was working, what percentage of shifts had a death occur?

3. Among the 1,384 shifts when Gilbert was not working, what percentage of shifts had a death occur?

4. Compute the difference between the percentage of shifts in which a death occurred when Gilbert was working and the percentage of shifts in which a death occurred when Gilbert was not working.

5. For this study, specify the explanatory variable and each of the possible categories of this variable.

6. For this study, specify the response variable and each of the possible response categories.

7. Were shifts that Gilbert was working more likely to have a death occur than on shifts when she was not?

8. Does the difference in percentages convince you that Gilbert was giving lethal injections of epinephrine to patients? Why or why not?

9. What might other possible explanations be for the difference between the two percentages?

Modeling the Chance Variation Under the Assumption of No Difference

You will conduct a randomization test using TinkerPlots™ to find out what differences in sample percentages you would see just by chance, assuming there is no difference between the percent of shifts in which a death occurred when Gilbert was working and those in which she was not working.

- Open the *Murderous-Nurse.tp* data set.
- Set up a sampler to run a randomization test. (If you have forgotten how, refer back to the *Contagious Yawns* activity for an example.)
- Carry out the randomization test with 500 trials and plot the 500 differences in percentages.

10. Sketch the plot below.

11. What are the cases in the plot? (Hint: ask yourself what each individual dot represents.)

12. Where is the plot of the results centered (at which value)? Explain why this makes sense based on the null hypothesis.

Evaluating the Hypothesized Model

13. Report the *p*-value (i.e., level of support for the hypothesized model) based on the observed result.

14. Based on the *p*-value, provide an answer to the research question.

15. Can we make cause-and-effect inferences and attribute the differences in death rate to the fact that Kristen Gilbert worked the shift? Explain based on the study design. If not, provide an alternative explanation for the difference in percentages.

16. Are the differences in death rate generalizable to the population of all 8-hour shifts at the hospital? Explain based on the study design.

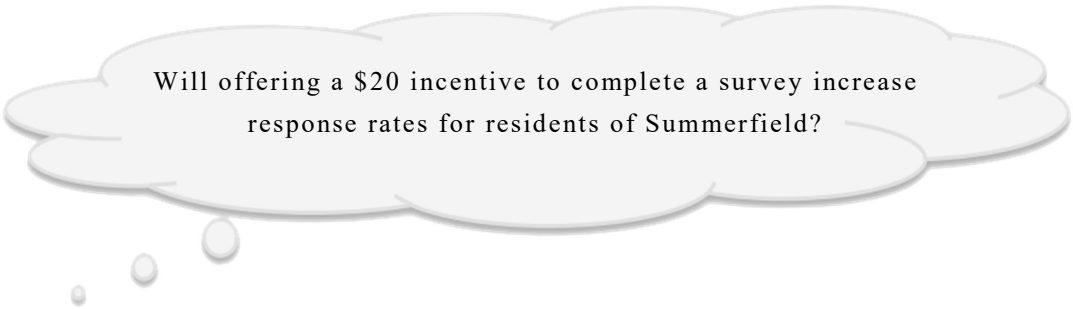
Appendix B6: Survey Incentives activity

Course Activity: Survey Incentives

Researchers who conduct surveys often have the problem of nonresponse. When response rates are low, it is hard to make valid conclusions from a survey, because people who respond may have different opinions from people who do not respond. One possible way to deal with this is to offer monetary incentives for responding. However, this can be costly, and if the incentive does not make it more likely that people will respond, then it is not worth spending the money.

In this activity we will consider the fictional town of Summerfield, which has 481 residents. The mayor of Summerfield wants to conduct a survey about the quality of life and improvements that could be made to the town, but is worried that many of the townspeople will not respond to the survey. She thinks it would be a good idea to offer survey respondents $20 to complete and return the survey. However, she does not want to spend a large amount of the town's budget on a financial incentive to respond if the incentive does not actually make people more likely to respond. Instead, she will first conduct a small pilot study to test the effectiveness of the survey incentive. You have been hired as a statistical consultant to help her design her study.

The mayor wants to answer the following research question:



Sampling

The mayor wants to conduct a study to see if people in Summerfield will be more likely to respond to a survey if they receive a $20 incentive than if they don't. The mayor wants to generalize her study results to the town, but she only has enough money to conduct a small pilot study with a maximum of 26 people (13 of whom would get the $20 incentive). The first step, therefore, is to choose who will be in the sample.

The first idea the mayor brings to you is to go door-to-door in her neighborhood and drop the survey into 26 mailboxes on or near her block.

1. How do you think these residents sampled from the mayor's neighborhood might differ from others in the town in their willingness to respond to the survey?

2. Should the mayor use this proposed sampling method? Explain why or why not.

Instead, the mayor has a list of all the adult residents of Summerfield in the town records.

3. How would you recommend she select her sample from this list? Be sure to provide her with enough detail that she can carry out this sampling method.

In addition to the list of residents, she has information from a recent town census on some of the population demographics regarding sex, age, income, and number of hours worked per week. Plots of the population demographics and parameters (population averages or percentages) are provided below.

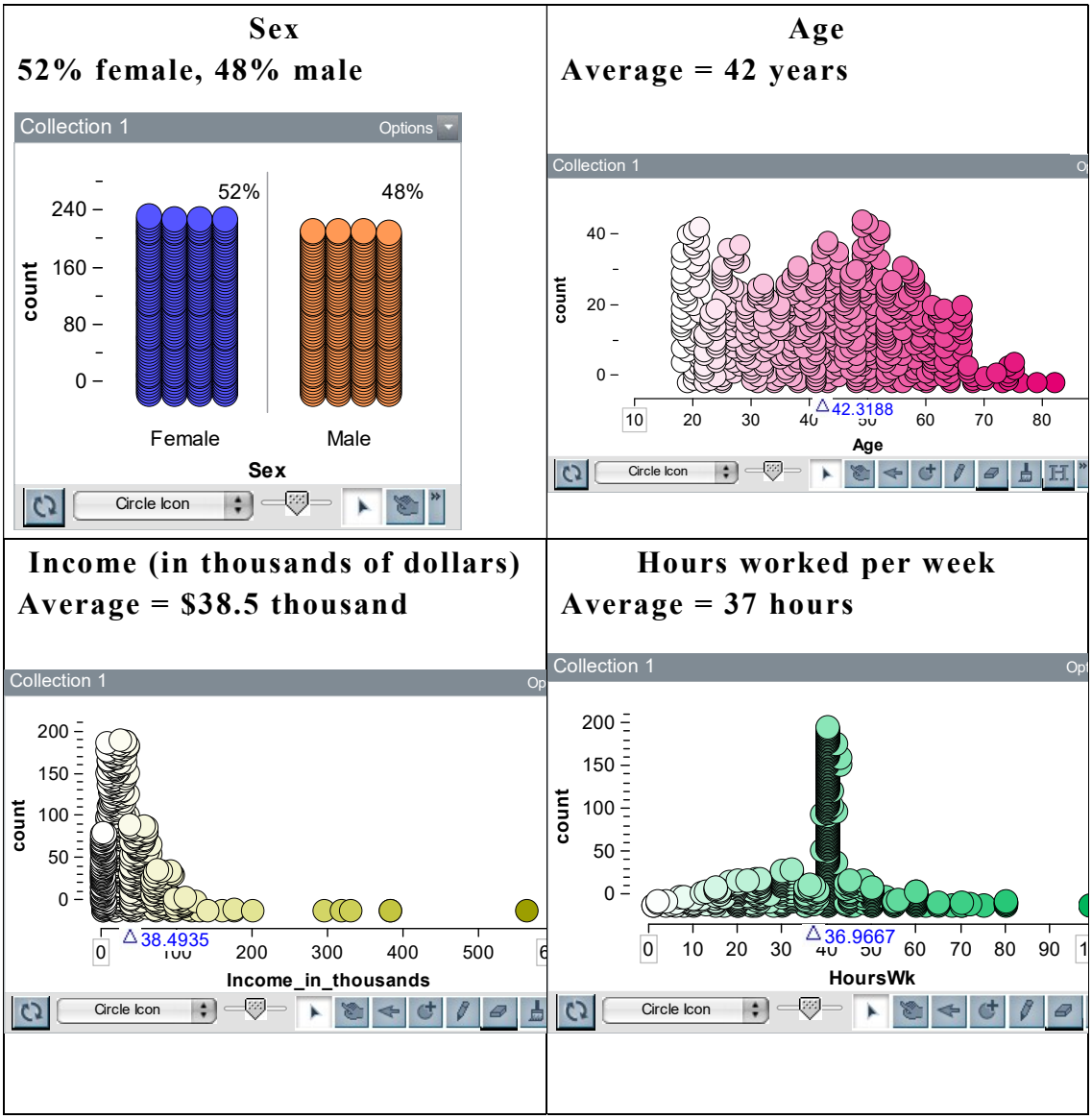


Figure 1. Population demographics of Summerfield.

You will now use TinkerPlots™ to simulate drawing a random sample from the population of Summerfield, and compare your sample demographics to the population. You will be plotting the variables sex, age, income, and hours worked per week for your sample.

4. How do you expect your plots of these four variables for your sample to compare to the plots in Figure 1? Explain.

Open the file *TownSampling.tp*

A sampler has been set up for you to draw a simple random sample of 26 people. Run the sampler.

- Plot each of the 4 variables from your sample. (You will have 4 different plots – one for each variable.)
- Display the percentages for the Sex variable.
- Display the averages for the Age, Income, and Hours Worked variables.

Keep all four plots open in your TinkerPlots window. You will now examine each variable individually:

5. What proportion of your sample is female? Is this close to the percentage of the population that is female?

6. What is the average age in your sample? Does the distribution of ages look similar to that of the ages in the population?

7. What is the average income in your sample? Does the distribution of incomes look similar to that of the incomes in the population?

8. What is the average hours worked per week in your sample? Does the distribution of hours worked per week look similar to that of the hours worked per week in the population?

9. With your four plots still open, click the Run button in the sampler a few times. For each new sample, look at your four distributions and descriptive statistics. Do you get the exact same distribution and numbers each time? Why or why not?

10. Choose **one** of the variables you plotted. Write the name of that variable here.

- Collect a statistic from that variable (either the % of females, or the average of any of the three quantitative variables).
- Collect that statistic for 199 additional samples.
- Plot the 200 statistics from the random samples and obtain the average.

11. Where is your plot centered?

12. Is the center of your plot near the population parameter for this variable (see plots in Figure 1 above)?

13. Based on your plot, does random sampling appear to be an unbiased method of selecting townspeople for the survey?

14. Explain to the mayor why your proposed method of random sampling is better than her proposed method of sampling people from her neighborhood.

As discussed in the *Sampling Countries* activity, random sampling is an *unbiased* sampling method. As you probably noticed, each time you took a random sample, the distributions of the variables did not look exactly the same as the population distributions, and your sample statistics were not always exactly the same as your population parameters. This is because of sampling variability: every time a sample is taken, there is variability and you will get different distributions and sample estimates.

Although there is variability with random sampling, we do not have *bias* – that is, we are not more likely to sample wealthier residents than poorer residents; we are not more likely to sample men than women, etc. Every adult in the town has a fair chance of being in the sample. Random sampling is an *unbiased* sampling method. This means that statistics obtained using this method will not tend to be systematically higher or lower than the parameters – or "truth" – about the population.

Assignment to Groups

The mayor decides to follow your advice and take a random sample of 26 people from the town list., Next, she must think about how to assign the subjects into two groups: the incentive group (those who will receive the $20 incentive) and the control group (those who will receive no financial incentive). One thing that might be of concern is *confounding* variables. Recall that confounding variables are variables not being manipulated by the researcher that can affect the results of the study.

Recall that we have access to information about four variables from the population. For the remainder of the activity, we will focus on only the three *quantitative* variables: age, income, and hours worked per week.

15. Which of these three variables do you think might be potential confounding variables that would affect residents' willingness to respond, regardless of whether or not they receive the incentive?

16. Explain how your confounding variable(s) of choice might affect the results of the mayor's study if she is not careful in how she assigns subjects to treatments.

Now, suppose the mayor has already taken a random sample of size 26. She then finds, however, that one of the people in the sample has very recently moved away. Therefore, she is left with a sample of size 25.

17. How would you advise her to assign the 25 subjects to the incentive and control groups? Be sure to provide her with enough detail that she can carry out this method.

One thing to note here is that even though we would ideally like to have equal sample sizes for the treatment and control groups, it is still all right to have two groups that are unequal in size. We can still compare two groups of unequal sizes because we can compare summary measures of the two groups, such as averages and proportions.

You will now use TinkerPlots™ to simulate randomly assigning 12 subjects to receive the survey with the $20 incentive (incentive group) and 13 subjects to receive the survey without the $20 incentive (control group).

- Open the file *TownAssignment.tp*

Note that the model has already been set up for you; there is a Counter device with the study participants and a Stacks device that is randomly assigning the group that participant will be in.

- Click Run to record the results of a single random assignment.

Choose one of the quantitative variables (age, income, or hours worked per week) that you think could be a potential confounding variable.

- Plot that variable on the *x*-axis and the Group variable on the *y*-axis.
- Obtain the average for each group.

18. Do the incentive and control groups appear similar to each other with respect to this confounding variable? Explain.

- Run the sampler a few more times and observe how the plot of differences changes.

19. Do you get the exact same randomization each time? Explain why or why not.

Now, just like in the *Strength Shoe* activity, for the variable you chose, collect the difference in averages from your randomization as follows:

- Use the Ruler tool to compute the difference in averages between the two groups. (Note: Subtract the Control group from the Incentive group.)
- Right-click on the difference in averages and select Collect Statistic.
- Collect 499 more trials.
- Plot the 500 differences.
- Organize and fully separate the results (no bin lines) for the plot.
- Show the Average (and its numeric value) on both plots.

20. Where is your plot centered?

21. Based on your answers to the previous question, does it appear that random assignment is an effective method for balancing out this confounding variable for the incentive and control groups?

Conclusions: Random Sampling vs. Random Assignment

While it is rare for studies to feasibly implement both random sampling and random assignment, the mayor's study design allows her to both randomly select a sample from the town's population, and randomly assign subjects in the sample to receive the survey either with the $20 incentive or without the incentive.

Suppose now that the mayor has carried out her study using both random sampling and random assignment. In addition, suppose that she has found that those who received the incentive were significantly more likely to respond to the survey than those who did not ($p < .01$).

22. Can the mayor generalize this finding to the population, and conclude that across the town's population, those who receive the $20 incentive should be more likely to respond than those who do not? If so, what part of her study design allows her to conclude this and why?

23. Can the mayor conclude that providing the $20 incentive was the cause of the higher response rates for the incentive group? If so, what part of her study design allows her to conclude this and why?

24. The mayor is having trouble distinguishing beween the role of randomness in choosing a sample and the role of randomness in assigning treatments. She tells you that as long as there is something random about her study, she can make generalizations to the population *and* conclude that the treatment variable was the cause of any observed differences in the response variable. Write a short report in which you explain to her the problem with her reasoning. In your report, compare what you did in the first part of this activity (Random **Sampling)** with what you did in the second part of this activity (Random **Assignment**). How is the role of randomness different in each case?

Appendix C: Activities: online versions (readings included as part of activities)

Appendix C1: Sampling Countries activity (online)

COURSE ACTIVITY: SAMPLING COUNTRIES

In this activity, you will compare different ways of taking samples of countries of the world from a population of countries.

1. Think of 20 countries that you believe are representative of the countries in the world (i.e., they resemble the collection of all countries of the world). Fill in the list of countries in the table below.

Country
1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
15.
16.
17.
18.
19.
20.

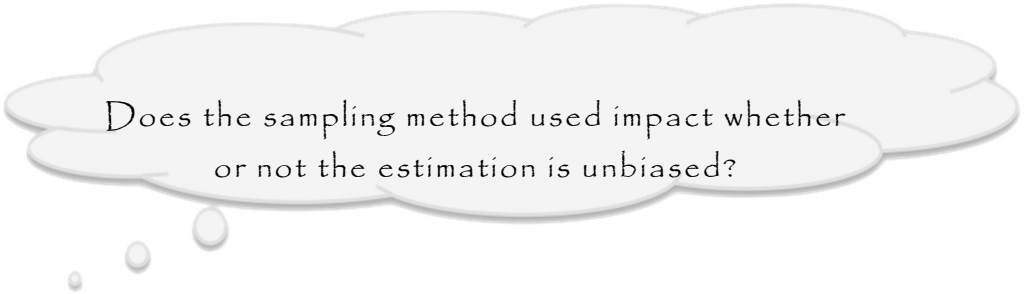
Group Question A:

a. Post in Moodle the list of the 20 countries you chose.
b. Describe how your list of countries is representative of the countries of the world.

In this activity, you will have access to a population of 196 countries of the world and some information about their life expectancy as determined by the World Bank (www.worldbank.org) in 2013. The data can be found in the last few pages of this activity. (Please note that not quite all of the countries of the world are in this dataset because some had missing data, but we will consider this list of 196 countries to be our *population* of countries.) You will examine the following variable of interest:

Life Expectancy: The number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.

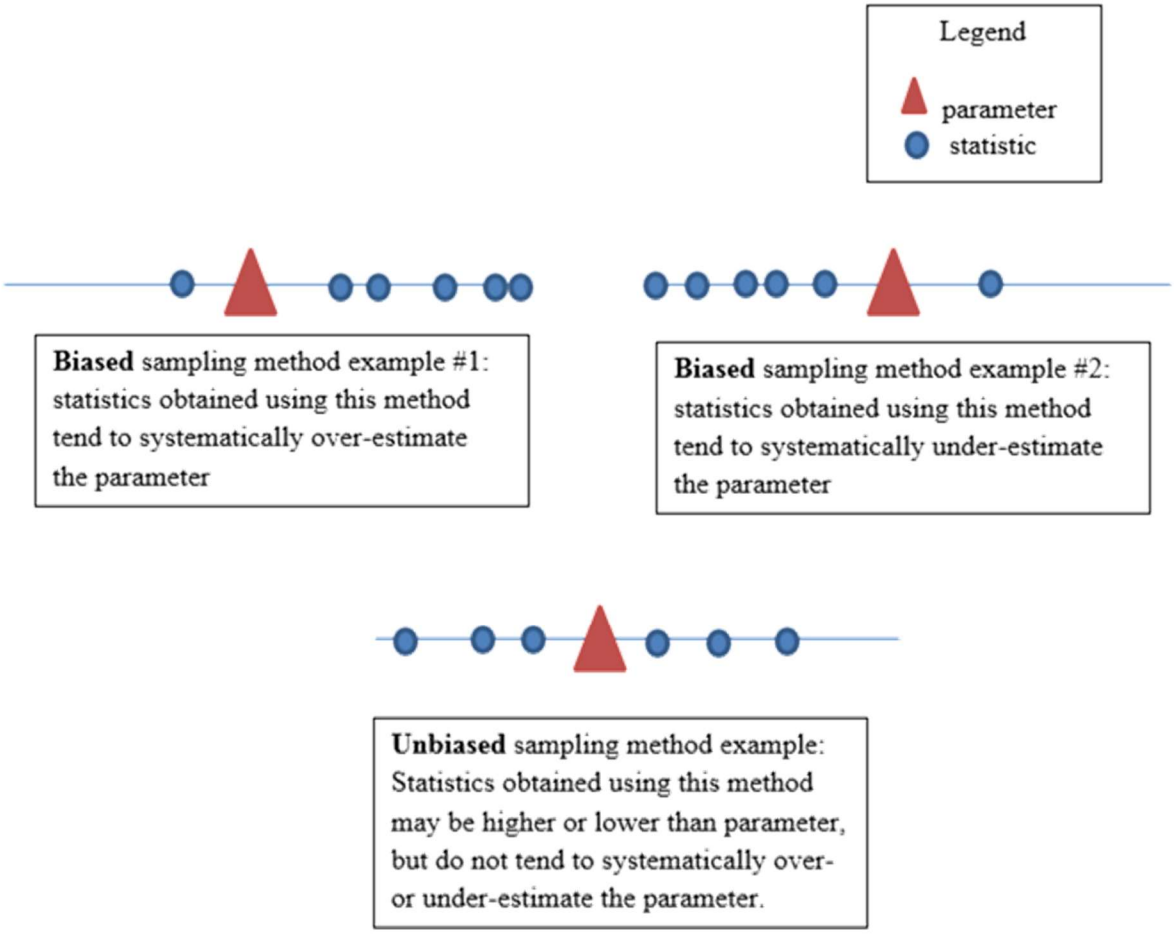
In this activity, you will be exploring the following research question:



UNBIASED ESTIMATION

One concern when taking a sample is whether or not an estimate taken from a sample (**statistic**) will appropriately estimate the "truth" of the population (**parameter**). When a sampling method produces statistics that tend to systematically over- or under-estimate the population parameter, we call that sampling method **biased.** Ideally, we want sample estimates to be **unbiased**. Unbiasedness means that the estimation method used tends to produce sample statistics that are around the population parameter, without the tendency to over-estimate or under-estimate the parameter.

For example, as illustrated in the picture below, suppose we are trying to estimate a parameter of the population, symbolized by a triangle. Statistics taken from different samples will vary, as symbolized by the small circles. The biased sampling method examples show how biased methods produce estimates that that tend to be higher or lower than the parameter we are trying to estimate. In contrast, the unbiased sampling method example shows how some estimates are on the low side, some estimates are on the high side, but as a whole they are centered on the true value of the parameter.



In statistics, **estimation** refers to the process by which one makes inferences about a population or model, based on information obtained from a sample. The **population** is the entire collection of who or what (e.g., the observational units) that you would like to draw inferences about. In practice, it is often impossible to examine every unit of the population, so data from a subset, or **sample**, of the population is examined instead. The sample data provides statisticians with the best estimate of the exact "truth" about the population. The "truth" one is searching for in the population is typically a summary measure such as the population average or population percentage. Summary measures of a population are called **parameters**. The estimates of these values from sample data are referred to as **statistics**

Follow these instructions to compute and report the average life expectancy for your sample of countries:

- Open up a blank TinkerPlots™ file.
- Drag a Table from the Object toolbar into your document.
- Create a new attribute called *Life Expectancy* in the first column of the case table.
- Using the tables at the back of this activity for reference, enter the life expectancies of your 20 countries under the *Life Expectancy Column*.
- Plot the 20 life expectancies.
- Separate and stack the cases, then find the value of the Average.

2. Write down the value of the average life expectancy of your 20 countries here.

3. Is this value a parameter, or a statistic?

- Open the file *Countries-Hand-Picked.tp*, available in Moodle. The table and plot show the mean life expectancies that were collected from 14 other hand-picked samples of 20 countries each.
- In the table "Hand Picked Sample Means," in a new row at the bottom, type the average (mean) life expectancy for your 20 countries.
- On your plot, place a vertical reference line at the value 71, which is the value of the population average life expectancy of all 196 countries.
- Click on the row number (15) next to the value you just added. This will highlight your new value on the plot.

Group Question B:

a. Paste into Moodle a copy of your plot of the sample average life expectancies.
b. Were most of your sample estimates around the population average of 71 years?
c. Approximately what percentage of these hand-picked sample means had sample statistics that exceeded the population average?

Group Question C:

a. Based on your answers to Group Question B, does this method of sampling appear unbiased, or does it tend to over-estimate or under-estimate the average life expectancy of the population of countries?
b. What are some reasons for why the sampling method of asking people to name 20 countries might produce biased estimates?

In order to try to eliminate potential biases that can occur by human selection, it is better to take a **random sample**. Humans are not very good "random samplers" – even though we are trying to obtain a representative sample, we tend to name countries that are more well known or appear more often in the news than others. Instead, it is important to use random sampling techniques to do the sampling for us.

The goal of random sampling is to obtain a representative sample, so estimates of population parameters are unbiased. Although there is variation from sample to sample, there is no systematic tendency to over-estimate, or to under-estimate, the population parameter.

SIMPLE RANDOM SAMPLING

A **simple random sample** (**SRS**) is a specific type of random sample that gives every observational unit in the population the same chance of being selected. In fact, every sample of size n has the same chance of being selected. In this example, we will take a simple random sample of 10 countries.

The first step in drawing a simple random sample is to obtain a **sampling frame**, which is a list of each member of the population (in this case, this will be a list of all of the countries in our population). We have already prepared a sampling frame of the countries for you.

USE TINKERPLOTS™ TO DRAW A SIMPLE RANDOM SAMPLE

- Download from Moodle the file *SamplingCountries.tp* and open it.
- Draw one simple random sample of 10 countries from the sampler. (Note that the sampler has been set up to draw the sample without replacement so you do not get any duplicates.)

First, you will examine the distribution of life expectancy for this single sample.

4. Plot the "Life Expectancy" variable for this single sample.

5. Obtain and record the average life expectancy for this single sample.

6. Do you think you and your group members will all obtain the exact same plot and sample average? If not, do you think you will obtain similar plots and sample averages?

7. Now, compare your average from this sample to the true population average of 71 years. Are the averages the same? Are they similar?

You may notice that your sample will differ from other samples taken by your classmates. Samples differ, but hopefully, your sample estimate should be somewhat close to the population average life expectancy, if it is a representative sample.

Now, we will investigate whether random sampling produces sample estimates that are unbiased.

8. If we took many *random* samples of size 10 and made plots of the sample average life expectancies similar to your plot in Group Question B, what do you think this plot would look like?

9. Where do you predict this plot will be centered?

In TinkerPlots™, go back to the plot of the life expectancies from the sample of size 10 you just examined.

- Collect the average life expectancy from your random sample.
- Carry out 200 trials of the simulation.
- Plot the 200 average life expectancies you collected.
- Obtain the average from your plot of the 200 sample statistics.

Group Question D:

a. Paste into Moodle a copy of your plot of the 200 samples.
b. If the sampling method is unbiased, where should you expect the plot to be centered? Is your plot centered near that value?
c. Based on your answer to the previous question, does simple random sampling produce an unbiased estimate of the average country's life expectancy? Explain.

Group Question E:

a. Compare your plot above in Group Question D with the plot you made in Group Question B. What do you think is better: taking a larger convenience sample (n = 20), or taking a smaller, random sample (n = 10)? Explain your choice.
b. When you draw a single random sample from a population, do you expect your sample statistic to match the population parameter exactly? Why or why not?
c. What does it mean for a sampling method to be unbiased?

Because random sampling is an unbiased sampling method, it allows us to use our samples to make generalizations, or wider inferences, about the population from which the sample was taken.

In real studies, researchers do not have access to information about the full population like you did in this activity. However, they need to use a sampling method that tends to produce representative samples that give unbiased estimates, so that they can make valid generalizations to the population of interest. For example, if a researcher took a random sample of countries from this population and found the sample average life expectancy to be 72.5, (s)he could generalize that the average country's life expectancy from this population is approximately around 72.5.

Country Name	Life Expectancy
Afghanistan	60.03
Albania	77.54
Algeria	74.57
Angola	51.87
Antigua and Barbuda	75.78
Argentina	75.99
Armenia	74.56
Aruba	75.33
Australia	82.20
Austria	80.89
Azerbaijan	70.69
Bahamas, The	75.07
Bahrain	76.55
Bangladesh	71.25
Barbados	75.33
Belarus	72.47
Belgium	80.39
Belize	69.98
Benin	59.31
Bermuda	80.57
Bhutan	69.10
Bolivia	67.91
Bosnia and Herzegovina	76.28
Botswana	64.36
Brazil	74.12
Brunei Darussalam	78.55
Bulgaria	74.47
Burkina Faso	58.24
Burundi	56.25
Cabo Verde	72.97
Cambodia	67.77
Cameroon	55.04
Canada	81.40
Central African Republic	49.88
Chad	51.19
Channel Islands	80.46
Chile	81.20

Country Name	Life Expectancy
China	75.35
Colombia	73.81
Comoros	62.93
Congo, Dem. Rep.	58.27
Congo, Rep.	61.67
Costa Rica	79.23
Cote d'Ivoire	51.21
Croatia	77.13
Cuba	79.26
Cyprus	79.95
Czech Republic	78.28
Denmark	80.30
Djibouti	61.69
Dominican Republic	73.32
Ecuador	75.65
Egypt, Arab Rep.	70.93
El Salvador	72.50
Equatorial Guinea	57.29
Eritrea	63.18
Estonia	76.42
Ethiopia	63.44
Faeroe Islands	81.30
Fiji	69.92
Finland	80.83
France	81.97
French Polynesia	76.33
Gabon	63.84
Gambia, The	60.00
Georgia	74.08
Germany	81.04
Ghana	61.14
Greece	80.63
Grenada	73.19
Guam	78.87
Guatemala	71.49
Guinea	58.22
Guinea-Bissau	54.84
Guyana	66.31
Haiti	62.40

Country Name	Life Expectancy
Honduras	72.94
Hong Kong	83.83
Hungary	75.27
Iceland	83.12
India	67.66
Indonesia	68.70
Iran, Islamic Rep.	75.13
Iraq	69.47
Ireland	81.04
Israel	82.06
Italy	82.29
Jamaica	73.47
Japan	83.33
Jordan	73.90
Kazakhstan	70.45
Kenya	60.95
Kiribati	65.77
Korea, Dem. Rep.	69.79
Korea, Rep.	81.46
Kuwait	74.46
Kyrgyz Republic	70.20
Lao PDR	65.69
Latvia	73.98
Lebanon	80.13
Lesotho	49.33
Liberia	60.52
Libya	71.66
Liechtenstein	82.38
Lithuania	74.16
Luxembourg	81.80
Macao SAR, China	80.34
Macedonia, FYR	75.19
Madagascar	64.67
Malawi	61.47
Malaysia	74.57
Maldives	76.60
Mali	57.54
Malta	80.75
Mauritania	62.80

Country Name	Life Expectancy
Mauritius	74.46
Mexico	76.53
Micronesia, Fed. Sts.	68.97
Moldova	68.81
Mongolia	69.06
Montenegro	74.76
Morocco	73.71
Mozambique	54.64
Myanmar	65.65
Namibia	64.34
Nepal	69.22
Netherlands	81.10
New Caledonia	77.12
New Zealand	81.41
Nicaragua	74.51
Niger	60.83
Nigeria	52.44
Norway	81.45
Oman	76.84
Pakistan	65.96
Panama	77.42
Papua New Guinea	62.45
Paraguay	72.80
Peru	74.28
Philippines	68.13
Poland	76.85
Portugal	80.37
Puerto Rico	78.71
Qatar	78.42
Romania	74.46
Russian Federation	71.07
Rwanda	63.39
Samoa	73.25
Sao Tome and Principe	66.26
Saudi Arabia	74.18
Senegal	65.88
Serbia	75.14
Seychelles	74.23

Country Name	Life Expectancy
Sierra Leone	50.36
Singapore	82.35
Slovak Republic	76.26
Slovenia	80.28
Solomon Islands	67.72
Somalia	55.02
South Africa	56.74
South Sudan	55.22
Spain	82.43
Sri Lanka	74.24
St. Lucia	74.91
St. Vincent and the Grenadines	72.81
Sudan	63.17
Suriname	70.99
Swaziland	48.94
Sweden	81.70
Switzerland	82.75
Syrian Arab Republic	74.72
Tajikistan	69.40
Tanzania	64.29
Thailand	74.25

Country Name	Life Expectancy
Timor-Leste	67.52
Togo	59.13
Tonga	72.64
Trinidad and Tobago	70.31
Tunisia	73.65
Turkey	75.18
Turkmenistan	65.46
Uganda	57.77
Ukraine	71.16
United Arab Emirates	77.20
United Kingdom	80.96
United States	78.84
Uruguay	76.84
Uzbekistan	68.23
Vanuatu	71.67
Venezuela, RB	74.07
Vietnam	75.76
Virgin Islands (U.S.)	79.62
West Bank and Gaza	73.20
Yemen, Rep.	63.58
Zambia	59.24
Zimbabwe	55.63

Appendix C2: Strength Shoe activity (online)

COURSE ACTIVITY: STRENGTH SHOE®

ESTABLISHING CAUSATION

Researchers often examine relationships between variables. Two variables are associated if the values of one variable tend to be related to the values of another variable. In particular, an **explanatory variable** is a variable that can be used to help us understand or predict values of the **response variable**.

In many studies, the goal is more than to determine an association. The goal is to determine whether changes in an explanatory variable influence, or cause, changes in a response variable. However, association does not necessarily mean that there is a **cause-and-effect** relationship: namely, that changing the values of one variable will influence the value of another variable. Consider this example:

Suppose educators are trying to figure out if taking a test preparation class will increase students' test scores. Students are allowed to choose whether to take the class or not, and in the end, the data show that the students who took the class scored significantly higher on the test than the students who did not ($p < .05$).

Here, the explanatory variable is whether or not the students took the class, and the response being measured is the test score. The researchers found a significant association between these variables.

But can the researchers conclude that the test preparation class was effective? Not necessarily. Think about how students who chose to take the class might be different from students who chose not to take it. Perhaps the students who chose to take the class would have had higher scores even if they had not taken the class, just because they're already more motivated to succeed or have higher GPAs than students who did not take the class. In this case, students' motivation and GPA are called **confounding variables** because they help to offer a plausible explanation for the observed association.

A study where researchers do not manipulate the explanatory variable is called an **observational study.** In this type of study, researchers may compare groups, but do not control which group a participant is in. The exam preparation class scenario above is a good example of an observational study, because the subjects choose whether or not to take the class. The researcher did not control this. The problem with observational studies is that cause-and-effect conclusions are difficult to make because the groups of

participants being compared may differ in ways other than the explanatory variable, and confounding can come into play.

In contrast, in an **experiment**, the researcher actively has control over which group each subject is in. When categories of the explanatory variable are assigned to subjects in an experiment, the explanatory variable is also called a **treatment variable. (**Recall that you have already seen examples of treatment variables in course activities such as *Memorization* and *Sleep Deprivation*.)

Consider the above example of the test preparation class. If you were to assign students to take the class or not, how would you do this? It's important to try to make sure that students who are more motivated, have higher GPAs, or study longer, are *not* more likely to end up in one group than the other. If the students in the class were similar in all respects to the students who did not take the class, then if we found that students who took the class did significantly better on the test, we could argue this was because of the class. Since the only major difference between the groups is that one took the class and one did not, we can argue that the class led to the higher scores.

As we will see in the next activity, **random assignment** is a method to create groups that are similar in all respects except for the treatment imposed. Random assignment will not produce groups that are *exactly* equivalent to each other with respect to *every* possible confounding variable. However, assigning randomly means that subjects with certain characteristics will *not* be more likely to be in one group than the other. The goal is to create similar groups, so we can argue that any observed significant differences in the response variable are because of the only major difference between the groups: the treatment variable. Therefore, using random assignment has the potential to allow researchers to establish a *cause-and-effect* relationship between the explanatory and response variables.

STRENGTH SHOE®

The Strength Shoe® is a modified athletic shoe with a 4-cm platform attached to the front half of the sole. Its manufacturer claims that people who wear this shoe can jump farther than people who wear ordinary training shoes. In this activity you will be examining the following question:

How can you design a study to evaluate whether the manufacturer's claim about the Strength Shoe®

ANSWER THE FOLLOWING QUESTIONS

Suppose that you take a random sample of individuals by randomly selecting them from the population. You observe who does and does not wear the Strength Shoe®, and then compare the two groups' jumping ability.

1. Why would it be advantageous to take a random sample of individuals for this study?

Suppose you find in this study that on average, the group who wears strength shoes can jump much farther than the group who wears ordinary training shoes.

2. Do you think this is compelling evidence that strength shoes really increase jumping ability? Explain.

The problem with the evidence from the situation described above question#1 is that you do not know if whether or not someone wears the Strength Shoe® is the only way in which the two groups differ. The random sampling may allow you to generalize that within the wider population from which the sample was taken, people who wear Strength Shoes® are able to jump farther than those who wear ordinary training shoes. But we do not know if the Strength Shoes® were actually the *cause* of the improved jumping ability. For example, subjects who choose to wear the Strength Shoe® could be more athletic to begin with than those who opt to wear the ordinary training shoes, and this is why they can jump farther.

When researchers want to find if a treatment variable *causes* changes in a response, they control the treatment variable. They do this by assigning study participants to groups. One group may receive one treatment (e.g., jump with Strength Shoes®), and the other group may receive a comparison treatment (e.g., jump with ordinary training shoes). What we want to see is that both groups are approximately equal in terms of other variables, such as athletic ability, height, sex, etc. so that these other variables are *not* causing differences in jumping ability. We want to be able to say that the type of shoe is what is affecting jumping ability.

A 1993 study published in the *American Journal of Sports Medicine* investigated the Strength Shoe® claim using 12 intercollegiate track and field athletes as study participants[4]. Suppose you also want to investigate this claim, and you recruit 12 of your friends to serve as subjects. You plan to have six people wear a Strength Shoe® and the other six wear the ordinary training shoes they usually wear when exercising, and then measure each group's jumping ability.

CONFOUNDING VARIABLES

Two factors that might affect jumping distance are a person's sex and their height. In every study, there are potentially many factors (aside from the treatment) that may be related to the response variable and, in turn, affect the results of the study. Statisticians refer to these variables as **confounding variables**.

One potential way to balance out some confounding variables would be to purposefully try to balance out certain confounding variables and create two groups that are relatively equivalent with respect to known confounding variables. Suppose the researcher decided to control for sex and height by purposefully assigning the two groups so that there was an equivalent number of females in each group, and the average height for each group was roughly equivalent:

[4] Cook, S. D., Schultz, G., Omey, M. L., Wolf, M. W., & Brunet, M. F. (1993). Development of lower leg strength and flexibility with the strength shoe. *American Journal of Sports Medicine, 21*, 445–448.

Ordinary Training Shoe Group		
Name	Sex	Height
Jasmine	Female	61
Ka Nong	Female	67
George	Male	67
Paul	Male	73
Tong	Male	71
Ringo	Male	71

Strength Shoe Group		
Name	Sex	Height
Keyaina	Female	63
Mary	Female	66
Antonio	Male	68
Andreas	Male	70
Davieon	Male	70
John	Male	69

Now, you will compare the two groups in the above sample based on height.

- Open the file *StrengthShoe-Purposeful.tp* found on Moodle.

Next, plot the height variable as follows:

- Plot the attributes Height (*x*-axis) and Group (*y*-axis) in a single plot.
- Separate and stack the cases.
- Display the average for each group.

3. Examine your plot of heights and write down the average height for each of the two groups. Also, compute the difference in the two averages.
   Average Height of Strength Shoe Group: _____
   Average Height of Ordinary Training Shoe Group: _____

Difference in average height (Strength Shoe® – Ordinary Training Shoe):

If the two groups are balanced with respect to sex and height, then if you find a significant difference in jumping ability between the two groups, you can argue that it was not differences in sex or height that caused the difference in jumping ability.

Group Question A:

a. Are the two groups roughly equivalent with respect to height?
b. Based on the tables on the previous page, does it appear that the two groups are equivalent to each other with respect to the Sex variable? Explain.
c. Can you think of other variables besides sex and height that might explain differences in jumping ability? If so, what are they?

Suppose now that there is a genetic factor (which you did not measure before the study) that will strongly influence how far participants will be able to jump (regardless of the shoe type). Let's call it the "*X*-factor." Since you do not know about it, you have no way to measure and control for it, but it will likely influence the results of our study. For

example, what if more participants assigned to the Strength Shoe® group have this *X*-factor? Then the Strength Shoe® group would show increased jumping ability, even if training with a StrengthShoe® is no better than training with an ordinary training shoe.

- To explore this, we actually have an *X*-Factor already hidden in the TinkerPlots™ file! To show it, right click anywhere in the table you see in the TinkerPlots™ window and select Show Hidden Attribute.

You will now see that the *X*-Factor variable ("Yes" or "No", indicating whether the participant has that genetic factor or not) has appeared as another attribute in the trial results. It is important to remember that in real life, you would not know about this confounding variable. But, here, you can examine how this confounding variable is distributed between the Strength Shoe® and Ordinary Training Shoe groups when the conditions are purposefully assigned.

- Plot the attributes X-Factor (*x*-axis) and Group (*y*-axis) in a single plot.
- Display the percentages for each group.

4. Calculate the percent of the subjects in each group that have the *X*-factor. Do you think the two groups are roughly equivalent with respect to whether or not they have the *X*-factor? Explain.

   Percent of people with *X*-factor in Strength Shoe Group: _____

   Percent of people with *X*-factor in Ordinary Training Shoe Group: _____

Difference in percent with *X*-factor (Strength Shoe® – Ordinary Training Shoe):

Suppose the 12 subjects were purposefully assigned to control for sex and height, as above. Researchers find that the subjects wearing the Strength Shoes® jump significantly farther, on average, than the subjects wearing ordinary training shoes.

Group Question B:

a. Record the difference in percent of subjects with the *X*-factor (Strength Shoe®– Ordinary Training Shoe).
b. Do you think we could conclude that the shoes caused the difference in jumping ability?

While some confounding variables may be identified and controlled in a study, others may not be identified initially by the researcher, such as the *X*-factor in this example.

Erroneous results because of unobserved confounding variables are prevalent in every field. Even the smartest and most experienced researchers will probably not identify all of the confounding factors related to differences in the response variable that need to be controlled.

Luckily, it turns out that the key to controlling for *all* of these confounding variables (both observed and unobserved) is to use *random assignment* in forming experimental groups. For the remainder of this course activity, you will examine how random assignment "equalizes" not only the observed confounding variables (e.g., height), but also unobserved confounding variables, like the *X*-factor.

RANDOM ASSIGNMENT

Random assignment is the preferred method of assigning subjects to treatment conditions in an experiment. Under random assignment, each subject has an equal chance (probability) of being assigned to any of the treatment conditions.

OBSERVED VARIABLE: HEIGHT

Next you will use TinkerPlots™ to randomly assign subjects to the two groups. Then, we want you to examine the average height in each group to see if height differences in the two groups could explain the jumping differences we saw between the groups.

- Open the *Strength-Shoe-Random-1.tp* TinkerPlots™ file found on Moodle.

Note that the model has already been set up for you; there is a Counter device with the study participants and a Stacks device that will randomly assign each participant to a group.

- Press Run to record the results of a single random assignment.
- Plot the attributes Height (*y*-axis) and Group (*x*-axis) in a single plot;
- Separate and stack the cases.
- Display the average for each group.

5. Calculate the average height for each group. Also find the difference in these two averages (taking the Strength Shoe® group's average minus the ordinary training shoe group's average).

   Average height in Strength Shoe® Group: _____

   Average height in Ordinary Training Shoe Group: _____

Difference in average height (Strength Shoe® – Ordinary Training Shoe):

Group Question C:

a. In this single random assignment, are the two groups exactly balanced with respect to height? Explain.
b. This is just a single random assignment, and we want to get a sense of the difference in the average height across many random assignments. Suppose you want to make a plot of the difference in average height for many different random assignments. Where do you predict this plot will be centered?

Now, construct this plot as follows:

- Use the Ruler tool to compute the difference in the average heights between the two groups. (Note: Subtract the Ordinary Training Shoe group from the Strength Shoe® group.)
- Right click on the difference in averages and select Collect Statistic.
- Collect 499 more trials.
- Plot the 500 differences.
- Organize and fully separate the results (no bin lines) for the plot.
- Show the Average (and its numeric value) on the plot.

Group Question D:

a. Paste into Moodle a copy of your plot of these 500 differences in average heights.
b. Where is this plot centered?
c. What does your answer to the previous question imply about the tendency of random assignment to balance out the height variable in the two groups? Explain.

UNOBSERVED CONFOUNDING VARIABLE

As we saw earlier, the variable *X*-Factor is an unobserved confounding variable (or lurking variable) that the researcher does not observe, but will strongly influence how far participants can jump, regardless of the shoe they use.

- Save your TinkerPlots™ file for future reference.
- Open the *Strength-Shoe-Random-2.tp* TinkerPlots™ file found on Moodle and Run the Sampler.

- Again, we actually have an *X*-Factor already hidden in the TinkerPlots™ file! To show it, right click anywhere in the table of trial results and select Show Hidden Attribute.
- Plot the attributes X-Factor (*x*-axis) and Group (*y*-axis) in a single plot.
- Organize and separate the cases based on both attributes.
- Display the percentage for each group.

6. Calculate and report the percent of people with the *X*-Factor in each group. Also subtract these two percentages (subtract the ordinary training shoe group's percent from the Strength Shoe® group's percent).
   Percent of people with the *X*-Factor in Strength Shoe® Group: _____
   Percent of people with the *X*-Factor in Ordinary Training Shoe Group: _____

Difference in percentages of people with *X*-Factor (Strength Shoe® – Ordinary Training Shoe):

Again, this is just a single random assignment and we want to get a better sense of the differences in the *X*-Factor across many random assignments to groups.

7. Suppose you want to make a plot of the difference in percentages of people with the *X*-Factor across many random assignments. Where do you predict this plot will be centered?

We need to again collect *two measures*: the percentage of participants with the *X*-Factor in the Strength Shoe® group and the percentage of participants with the *X*-Factor in the Ordinary Training Shoe group.

Create a plot of the differences in percentage of people with the *X*-Factor for each group as follows:

- Right-click on the percent with the *X*-Factor for the Strength Shoe® group and select Collect Statistic.
- Right-click on the percent with the *X*-Factor for the Ordinary Training Shoe group and select Collect Statistic.
- Create a third attribute in your History of Results table and name it *Difference*.
- Right-click *Difference* and select Edit Formula.
- Use the Formula Editor to compute the difference in the percent of people with the *X*-Factor between the two groups. (Note: Subtract the Ordinary Training Shoe group from the Strength Shoe® group.)
- Collect 499 more trials.
- Plot the 500 differences.
- Organize and fully separate the results (no bin lines) for the plot.
- Show the Average (and its numeric value) on the plot.

Group Question E:

a. Paste into Moodle a copy of your plot of these 500 differences in percentages of people with the *X*-factor.
b. Where is this plot centered?
c. What does your answer to the previous question imply about the tendency of random assignment to balance out the *X*-factor variable in the two groups? Explain.

Group Question F:

a. Which method of assignment is better: Purposefully assigning groups to balance out known confounding variables, or assigning groups randomly? Explain.

b. Suppose we conduct this random assignment and find that the Strength Shoe® group jumps significantly farther, on average, than the ordinary training shoe group. Would you be comfortable concluding that Strength Shoes® caused the increased jumping distance? How would you argue that most likely no confounding variable was responsible for this difference in jumping distance?

c. Now, look back at how the actual sample of 12 subjects was collected back on page 5. Would you be comfortable generalizing the results of a study based on that sample to conclude that training with Strength Shoe® will increase jumping distance for all athletes? Explain.

Appendix C3: Murderous Nurse activity (online)

COURSE ACTIVITY: MURDEROUS NURSE

SCOPE OF INFERENCES

The inferences one can draw from a statistical study depend on how the study was designed. There are two types of inferences researchers may wish to draw from a study:

(1) generalization to a population and

(2) making cause-and-effect conclusions.

These are two distinct types of conclusions, and randomness plays a different role in the study design for each.

GENERALIZATION TO A POPULATION

In statistical studies, we often wish to draw a conclusion about a population of interest, using a sample drawn from that population. In other words, we wish to **generalize** our results back to our population of interest. To generalize means to make a claim about a wider population of interest, using a sample of data. In order to do this, we need a representative sample, or one that is similar in characteristics to the population. Consider this example:

> *The Physician's Health Study[5] was conducted in the 1980's to study whether or not taking a daily low-dose aspirin reduced the risk of heart attacks. The sample was gathered by initially sending out letters to recruit male physicians between the ages of 40 and 84 who lived in the United States and who were registered with the American Medical Association. Using a sample of over 30,000 willing and eligible physicians, an experiment was conducted and it was found that the subjects who took the low-dose aspirin were significantly less likely to suffer from heart attacks than those who took a placebo.*

Can we say that for all adults in the U.S., those who take aspirin daily are less likely to suffer from heart attacks than those who do not? This would be making a **generalization** to the population of U.S. adults.

---

[5] http://phs.bwh.harvard.edu/

Given how the sample was taken, there could be **bias** in estimating the difference in heart attack rates between adults who take daily aspirin and adults who do not. This is because the sample of U.S. male physicians is likely not representative of all U.S. adults. Females are not represented. It's quite possible that male physicians have diets, exercise routines, and other health habits that are different from those of the general adult population. It is difficult to reliably make any generalizations about aspirin and heart attacks to a wider population of U.S. adults when the sample is a biased representation of this population.

In order to avoid this bias, the best way to obtain a representative sample is **random sampling**. By using randomness to select subjects, we eliminate human bias that makes some units more likely to be in the sample than others. Instead, we give all units in the population an equal chance to be in the sample. By sampling randomly, we are likely to get a sample that looks more or less like a snapshot of the population. An **unbiased** sampling method like random sampling means that we will not have a tendency to over-estimate or under-estimate the parameter of interest. Then, we can use the sample to make generalizations about the population that it represents.

ESTABLISHING CAUSATION

Making a cause-and-effect conclusion is a separate goal that may be desired from a study. In order to make causal claims, a significant association must first be found between a treatment variable and a response variable. If the two variables are associated, we can conclude there is a relationship, but we cannot necessarily conclude that changes in the treatment variable will lead to changes in the response variable. However, when we establish **causation**, we claim that changes in the treatment variable influence the value of the response variable.

In the example above with the Physician's Health Study, an association was observed: those who took daily aspirin were less likely to suffer from a heart attack than those who took a placebo. But can we conclude that the aspirin was the *cause* of the lower heart attack rates?

It's important to first consider whether any **confounding** variables – that is, variables that may be associated with both the treatment and response variables – could be responsible for the observed association. If the physicians were allowed to self-select whether they took aspirin or not, it's possible that those who chose to take aspirin might have tended to have healthier lifestyles than those who did not take any aspirin. Then, it

could be the difference in tendency to live a healthy lifestyle – not the aspirin itself – which might have been responsible for the difference in heart attack rates.

However, in the Physician's Health Study on aspirin, the subjects were randomly assigned into two groups: one took aspirin and one took the placebo. With this random assignment, we are not more likely to get subjects with healthier lifestyles in one group than another – everyone has an equal chance to be in each group. The random assignment tends to balance out the groups with respect to all potential confounding variables. If the only major difference between the groups was that one took aspirin and the other one took a placebo, then any differences in heart attack rates observed can be attributed to the aspirin vs. placebo treatment. Therefore, the fact that those who took aspirin were less likely to develop a heart attack is evidence that the aspirin lowered the heart attack rates.

TWO TYPES OF INFERENCES

Note that *generalization to a population* and *establishing causation* are two different types of inferences. Randomness is a desired part of the study design for making each of these inferences, but the type of randomness we need for each inference should not be confused.

With generalization, we ask the question: "Can we use the results from this sample to make a broader claim about the population the sample was taken from?" To answer yes, we ideally need random sampling to enable a sample that is representative of the population. Random sampling is intended to reduce the differences between sample and population, so that we can generalize to this population.

With establishing causation, we ask the question: "Does changing one variable lead to a change in another variable?" To answer yes, we ideally need random assignment in order to balance out confounding variables between treatment groups so that they are similar in all respects except for the level of the treatment variable. Random assignment is intended to reduce the differences between the two groups due to factors that are *not* being manipulated in the experiment, so that if there are differences in the response variable, we can attribute these differences to the treatment variable.

It is possible to have random sampling, random assignment, both, or neither. These different types of study designs and the inferences you can make from them are summarized in the table below.

		Selection of Units	
		Random Sampling	**No Random Sampling**
Allocation of Units to Groups	**Random Assignment**	Can make a causal conclusion and can generalize conclusion to the population	Can make a causal conclusion but cannot generalize this conclusion to the population
	No Random Assignment	Can generalize to the population, but cannot make causal claims	Cannot generalize to the population, and cannot make causal claims either

In the case of the Physician's Health Study, we had random assignment, but not random sampling. We can claim that taking daily aspirin reduces the chance of heart attack, but this may only be true for men similar in characteristics to those in the sample. We cannot necessarily generalize this claim to all adults, or even all males ages 40-84. Males in the overall population may be different from these physicians in characteristics such as health and exercise habits, and thus may respond differently to aspirin than the physicians in this study.

Ideally, it would be great if we could have both random sampling and random assignment, so that we could make causal claims that we could generalize to the population. The reality in study design is that it is difficult and rare to have a study that uses both random sampling and random assignment. In order to perform an experiment with random assignment, often the study participants have to give up a lot of time. It is much easier to recruit people who are willing to give up their time and be in the study than to randomly sample from the population and rely on the people in this random sample to participate in the study.

In many cases, we have studies that have neither random sampling nor random assignment. With these studies, we cannot necessarily generalize findings to a population nor establish any causal claims. However, results from these studies may still reveal interesting findings and lead to further research.

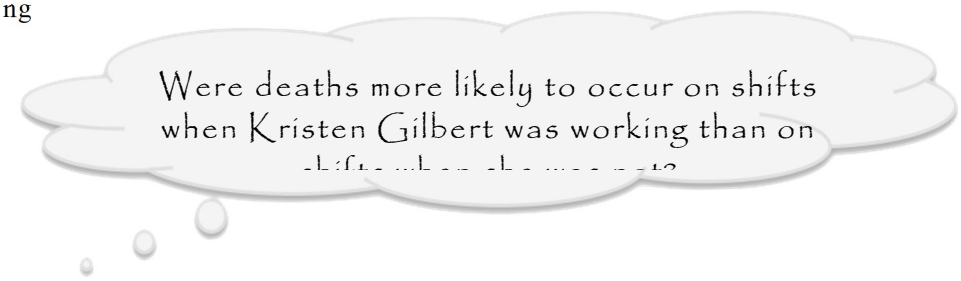
MURDEROUS NURSE

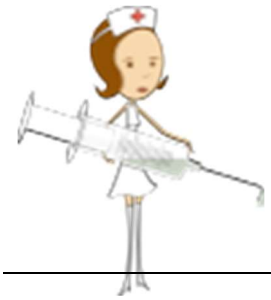
For several years in the 1990s, Kristen Gilbert worked as a nurse in the intensive care unit (ICU) of the Veteran's Administration hospital in Northampton, Massachusetts. Over the course of her time there, other nurses came to suspect that she was killing patients by injecting them with the heart stimulant epinephrine.

Part of the evidence against Gilbert was a statistical analysis of more than one thousand 8-hour shifts during the time Gilbert worked in the ICU[6]. Here are the data presented during her trial:

	Gilbert working on shift	Gilbert not working on Shift	Total
Death occurred on Shift	40	34	74
No death occurred on shift	217	1350	1567
Total	257	1384	1641

You will use these data to answer the following research question:





[6] Cobb, G. W., & Gehlbach, S. (2006). Statistics in the courtroom: United States vs. Kristen Gilbert. In R. Peck, G. Casella, G. Cobb, R. Hoerl, D. Nolan, R. Starbuck and H. Stern (Eds.), *Statistics: A guide to the unknown* (4th Edition), pp. 3–18. Duxbury: Belmont, CA.

ANSWER THE FOLLOWING QUESTIONS

1. Among all 1,641 shifts, what percentage of shifts had a death occur?

2. Among the 257 shifts when Gilbert was working, what percentage of shifts had a death occur?

3. Among the 1,384 shifts when Gilbert was not working, what percentage of shifts had a death occur?

Group Question A:

a. For this study, specify the explanatory variable and each of the possible categories of this variable.
b. For this study, specify the response variable and each of the possible response categories.

Group Question B:

a. Compute the difference between the percentage of shifts in which a death occurred when Gilbert was working and the percentage of shifts in which a death occurred when Gilbert was not working.
b. Were shifts that Gilbert was working more likely to have a death occur than on shifts when she was not?
c. Does the difference in percentages convince you that Gilbert was giving lethal injections of epinephrine to patients? Why or why not?
d. What might other possible explanations be for the difference between the two percentages?

MODELING THE CHANCE VARIATION UNDER THE ASSUMPTION OF NO DIFFERENCE

You will conduct a randomization test using TinkerPlots™ to find out what differences in sample percentages you would see just by chance, assuming there is no difference

between the percent of shifts in which a death occurred when Gilbert was working and those in which she was not working.

- Open the *Murderous-Nurse.tp* data set.
- Set up a sampler to run a randomization test. (If you have forgotten how, refer back to the *Contagious Yawns* activity for an example.)
- Carry out the randomization test with 500 trials and plot the 500 differences in percentages.

Group Question C:

a. Paste into Moodle your plot of these 500 differences in percentages.
b. What are the cases in the plot? (Hint: ask yourself what each individual dot represents.)
c. Where is the plot of the results centered (at which value)? Explain why this makes sense based on the null hypothesis.

EVALUATING THE HYPOTHESIZED MODEL

Group Question D:

a. Report the *p*-value (i.e., level of support for the hypothesized model) based on the observed result.
b. Based on the *p*-value, provide an answer to the research question.
c. Can we make cause-and-effect inferences and attribute the differences in death rate to the fact that Kristen Gilbert worked the shift? Explain based on the study design. If not, provide an alternative explanation for the difference in percentages.
d. Are the differences in death rate generalizable to the population of all 8-hour shifts at the hospital? Explain based on the study design.

Group Question E:

There are clearly limitations in this study. Do you believe that there is any value to examining observational data about the deaths that occur on Gilbert's shifts, compared to the deaths that occur on other shifts? Explain. Would it be advisable to conduct a

follow-up study where we randomly assign Kristin Gilbert to shifts in order to strengthen our inferences?

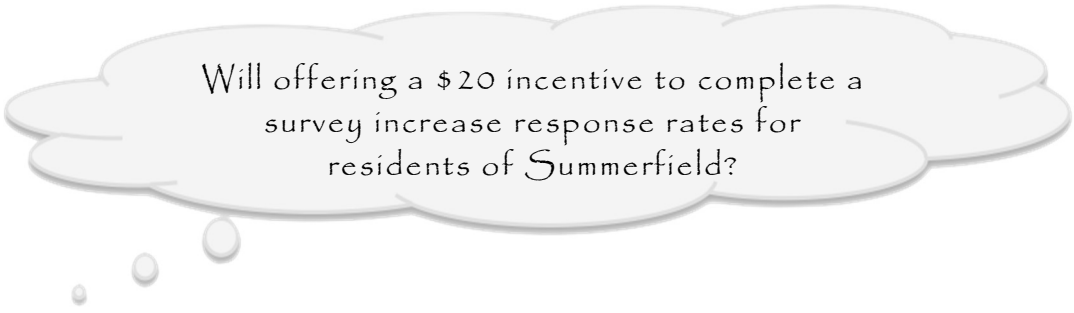
Appendix C4: Survey Incentives activity (online)

COURSE ACTIVITY: SURVEY INCENTIVES

Researchers who conduct surveys often have the problem of nonresponse. When response rates are low, it is hard to make valid conclusions from a survey, because people who respond may have different opinions from people who do not respond. One possible way to deal with this is to offer monetary incentives for responding. However, this can be costly, and if the incentive does not make it more likely that people will respond, then it is not worth spending the money.

In this activity we will consider the fictional town of Summerfield, which has 481 residents. The mayor of Summerfield wants to conduct a survey about the quality of life and improvements that could be made to the town, but is worried that many of the townspeople will not respond to the survey. She thinks it would be a good idea to offer survey respondents $20 to complete and return the survey. However, she does not want to spend a large amount of the town's budget on a financial incentive to respond if the incentive does not actually make people more likely to respond. Instead, she will first conduct a small pilot study to test the effectiveness of the survey incentive. You have been hired as a statistical consultant to help her design her study.

The mayor wants to answer the following research question:



SAMPLING

The mayor wants to conduct a study to see if people in Summerfield will be more likely to respond to a survey if they receive a $20 incentive than if they don't. The mayor wants to generalize her study results to the town, but she only has enough money to conduct a small pilot study with a maximum of 26 people (13 of whom would get the $20 incentive). The first step, therefore, is to choose who will be in the sample.

The first idea the mayor brings to you is to go door-to-door in her neighborhood and drop the survey into 26 mailboxes on or near her block.

Group Question A:

c. How do you think these residents sampled from the mayor's neighborhood might differ from others in the town in their willingness to respond to the survey?
d. Should the mayor use this proposed sampling method? Explain why or why not.
e. Instead, the mayor has a list of all the adult residents of Summerfield in the town records. How would you recommend she select her sample from this list? Be sure to provide her with enough detail that she can carry out this sampling method.

In addition to the list of residents, she has information from a recent town census on some of the population demographics regarding sex, age, income, and number of hours worked per week. Plots of the population demographics and parameters (population averages or percentages) are provided below.

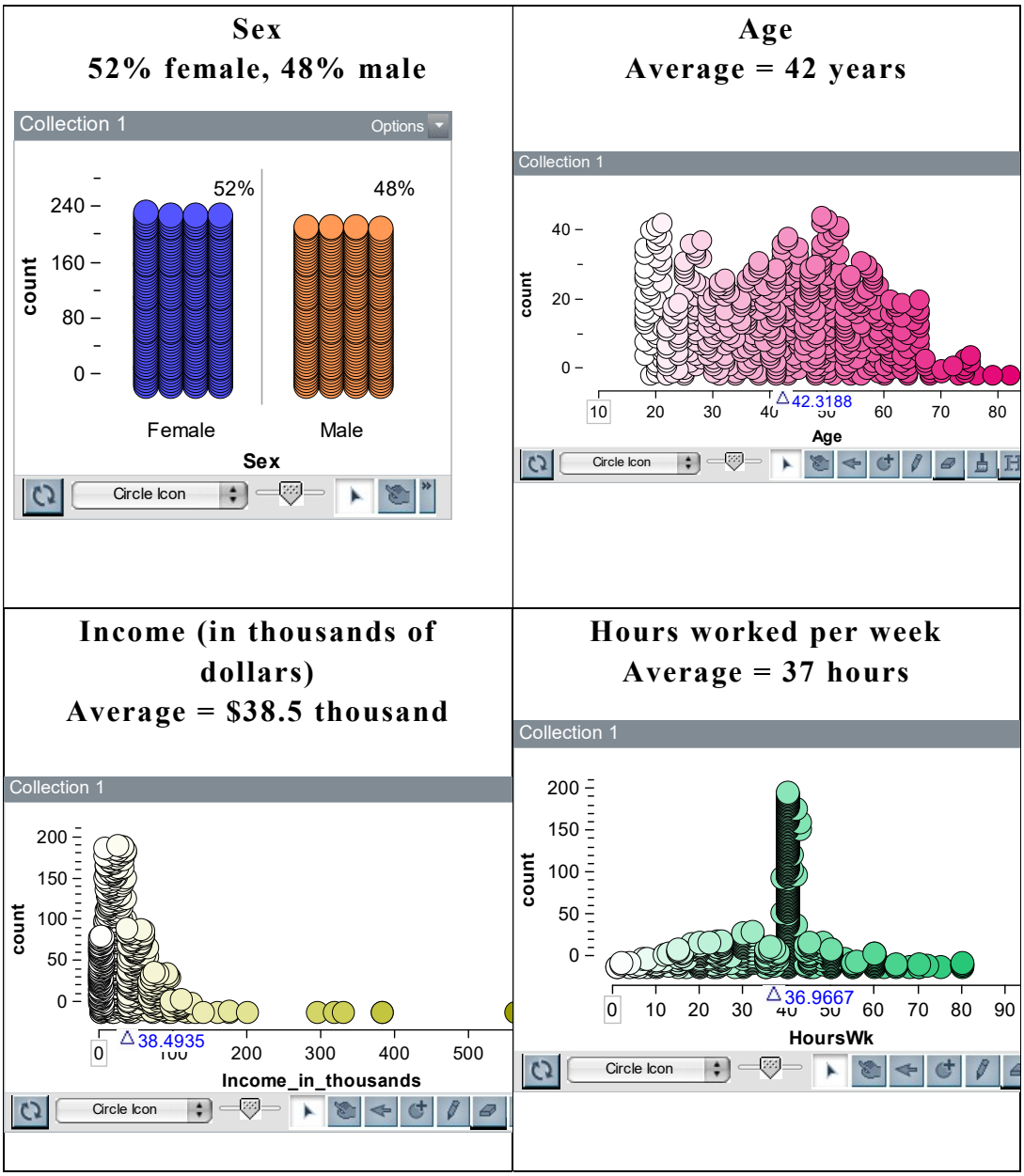


Figure 1. Population demographics of Summerfield.

You will now use TinkerPlots™ to simulate drawing a random sample from the population of Summerfield, and compare your sample demographics to the population. You will be plotting the variables sex, age, income, and hours worked per week for your sample.

1. How do you expect your plots of these four variables for your sample to compare to the plots in Figure 1? Explain.

- Open the file *TownSampling.tp*
- A sampler has been set up for you to draw a simple random sample of 26 people. Run the sampler.
- Plot each of the 4 variables from your sample. (You will have 4 different plots – one for each variable.)
- Display the percentages for the Sex variable.
- Display the averages for the Age, Income, and Hours Worked variables.

Keep all four plots open in your TinkerPlots window. You will now examine each variable individually:

2. What percentage of your sample is female? Is this close to the percentage of the population that is female?

3. What is the average age in your sample? Does the distribution of ages look similar to that of the ages in the population?

4. What is the average income in your sample? Does the distribution of incomes look similar to that of the incomes in the population?

5. What is the average hours worked per week in your sample? Does the distribution of hours worked per week look similar to that of the hours worked per week in the population?

6. With your four plots still open, click the Run button in the sampler a few times. For each new sample, look at your four distributions and descriptive statistics. Do you get the exact same distribution and numbers each time? Why or why not?

- Choose **one** of the variables you plotted. Write the name of that variable here.
- Collect a statistic from that variable (either the % of females, or the average of any of the three quantitative variables).
- Collect that statistic for 199 additional samples.
- Plot the 200 statistics from the random samples and obtain the average.

Group Question B:

a. Paste into Moodle a copy of your plot of the 200 statistics.
b. Where is your plot centered? Is the center of your plot near the population parameter for this variable (see plots in Figure 1 above)?
c. Based on your plot, does random sampling appear to be an unbiased method of selecting townspeople for the survey?
d. Explain to the mayor why your proposed method of random sampling is better than her proposed method of sampling people from her neighborhood.

As discussed in the *Sampling Countries* activity, random sampling is an *unbiased* sampling method. As you probably noticed, each time you took a random sample, the distributions of the variables did not look exactly the same as the population distributions, and your sample statistics were not always exactly the same as your population parameters. This is because of sampling variability: every time a sample is taken, there is variability and you will get different distributions and sample estimates.

Although there is variability with random sampling, we do not have *bias* – that is, we are not more likely to sample wealthier residents than poorer residents; we are not more likely to sample men than women, etc. Every adult in the town has a fair chance of being in the sample. Random sampling is an *unbiased* sampling method. This means that statistics obtained using this method will not tend to be systematically higher or lower than the parameters – or "truth" – about the population.

ASSIGNMENT TO GROUPS

The mayor decides to follow your advice and take a random sample of 26 people from the town list., Next, she must think about how to assign the subjects into two groups: the incentive group (those who will receive the $20 incentive) and the control group (those who will receive no financial incentive). One thing that might be of concern is *confounding* variables. Recall that confounding variables are variables not being manipulated by the researcher that can affect the results of the study.

Recall that we have access to information about four variables from the population. For the remainder of the activity, we will focus on only the three *quantitative* variables: age, income, and hours worked per week.

Group Question C:

a. Which of these three variables do you think might be potential confounding variables that would affect residents' willingness to respond, regardless of whether or not they receive the incentive?
b. Explain how your confounding variable(s) of choice might affect the results of the mayor's study if she is not careful in how she assigns subjects to treatments.
c. Now, suppose the mayor has already taken a random sample of size 26. She then finds, however, that one of the people in the sample has very recently moved away. Therefore, she is left with a sample of size 25. How would you advise her to assign the 25 subjects to the incentive and control groups? Be sure to provide her with enough detail that she can carry out this method.

One thing to note here is that even though we would ideally like to have equal sample sizes for the treatment and control groups, it is still all right to have two groups that are unequal in size. We can still compare two groups of unequal sizes because we can compare summary measures of the two groups, such as averages and proportions.

You will now use TinkerPlots™ to simulate randomly assigning 12 subjects to receive the survey with the $20 incentive (incentive group) and 13 subjects to receive the survey without the $20 incentive (control group).

- Open the file *TownAssignment.tp*

Note that the model has already been set up for you; there is a Counter device with the study participants and a Stacks device that is randomly assigning the group that participant will be in.

- Click Run to record the results of a single random assignment.

Choose one of the quantitative variables (age, income, or hours worked per week) that you think could be a potential confounding variable.

- Plot that variable on the *x*-axis and the Group variable on the *y*-axis.

- Obtain the average for each group.

7. Do the incentive and control groups appear similar to each other with respect to this confounding variable? Explain.

- Run the sampler a few more times and observe how the plot of differences changes.

8. Do you get the exact same randomization each time? Explain why or why not.

Now, just like in the *Strength Shoe* activity, for the variable you chose, collect the difference in averages from your randomization as follows:

- Use the Ruler tool to compute the difference in averages between the two groups. (Note: Subtract the Control group from the Incentive group.)
- Right-click on the difference in averages and select Collect Statistic.
- Collect 499 more trials.
- Plot the 500 differences.
- Organize and fully separate the results (no bin lines) for the plot.
- Show the Average (and its numeric value) on both plots.

Group Question D:

a. Paste into Moodle a copy of your plot of 500 differences.
b. Where is your plot centered?
c. Based on your answers to the previous question, does it appear that random assignment is an effective method for balancing out this confounding variable for the incentive and control groups?

CONCLUSIONS: RANDOM SAMPLING VS. RANDOM ASSIGNMENT

While it is rare for studies to feasibly implement both random sampling and random assignment, the mayor's study design allows her to both randomly select a sample from the town's population, and randomly assign subjects in the sample to receive the survey either with the $20 incentive or without the incentive.

Suppose now that the mayor has carried out her study using both random sampling and random assignment. In addition, suppose that she has found that those who received the incentive were significantly more likely to respond to the survey than those who did not ($p < .01$).

Group Question E:

a. Can the mayor generalize this finding to the population, and conclude that across the town's population, those who receive the $20 incentive should be more likely to respond than those who do not? If so, what part of her study design allows her to conclude this and why?
b. Can the mayor conclude that providing the $20 incentive was the cause of the higher response rates for the incentive group? If so, what part of her study design allows her to conclude this and why?

Group Question F:

The mayor is having trouble distinguishing beween the role of randomness in choosing a sample and the role of randomness in assigning treatments. She tells you that as long as there is something random about her study, she can make generalizations to the population *and* conclude that the treatment variable was the cause of any observed differences in the response variable. Write a short report in which you explain to her the problem with her reasoning. In your report, compare what you did in the first part of this activity (Random **Sampling)** with what you did in the second part of this activity (Random **Assignment**). How is the role of randomness different in each case?

Appendix D: Lesson plans for activities

Appendix D1: Sampling Countries lesson plan

Unit 3, Lesson 1

Sampling Countries

Summary

The *Sampling Countries* activity allows students to explore and compare different methods of sampling. It addresses the research question "Does the sampling method used impact whether the estimation is unbiased?" Students start with a brief discussion of how they could take samples of students from their class, and which methods might be better than others. Then, they move to the "Sampling Countries" portion of the activity, where the goal is to examine different sampling methods for estimating the average life expectancy. They first take convenience samples of size 20, and then random samples of size 10. The idea is to compare the different sampling methods, and explore how random sampling produces unbiased estimates. The sample sizes for the methods differ so that students can explore how a smaller, random sample is better than a larger, convenience sample because the method of random sampling is unbiased.

Learning Goals

This activity has the following goals for students:

- Understand the difference between biased and unbiased sampling methods
- Understand how human convenience sampling may lead to bias.
- Understand that random sampling produces estimates that are unbiased
- Understand that a smaller random sample is preferable to a larger, biased sample.

Reading Preparation

None. Students have taken the IDEA (Inferences from Design Assessment) as a part of Lab #7 as a pretest, without having had any reading background.

TinkerPlots files needed

SamplingCountries.tp

LESSON

To have on your computer before class
Have an open TinkerPlots table with an attribute called "Average Life Expectancy." Students will enter the sample average life expectancies from their convenience samples.

Begin the activity: students work together to pick what they believe to be a "representative" sample of countries. They obtain a sample average to plot on the instructor computer (~25 minutes)

Teacher Instructions: Briefly introduce the activity, say that we will be talking about methods of sampling from a population. Tell students they have approximately 20-25 minutes to get up to #4, and to stop when they get there. They will be giving you a value to enter into your computer on TinkerPlots.

Suggestions for potential issues:

Students may ask you what is meant by "representative." If so, you can ask:

- What set of 20 countries do you think might be a good snapshot of the collection of all countries of the world?

Once all students give you their value, STOP.

Teacher Instructions: Plot the values in the case table. When the plot is ready, tell students they can move on with the activity and then work through the end.

Students work through the rest of the activity (~30 minutes)

Suggestions for Potential Issues:

- I expect the students to name countries that are more easily recalled – even if they try to name countries from all continents, when I sorted the countries from highest to lowest life expectancy, I noticed many of the more "well-known" countries (e.g., the ones that appear in the news more often) have a higher life expectancy.

  But - IF it turns out your students are good representative samplers and the plot happens to be centered around 71, when you tell them the plot is ready to sketch, you can stop them for a brief large group discussion as you project the plot and ask them:

  - Where is this plot centered?
  - Where do you think this plot of sample averages *would* have been centered if I had asked you to name the first 20 countries you can recall, without asking you to make the sample "representative"?
    - Why?
  - Do you think this sampling method of naming the first 20 countries you can think of would have tended to over-estimate, or under-estimate the average life expectancy?

More Suggestions for Potential Issues:

- For the random sampling portion, students are taking random samples of size 10, and not 20 as they did before. If they ask why they are taking such small samples, mention that we are exploring the sampling *method*, and not to worry about sample size for now. (They will later answer a question about whether it's better to take a smaller random sample or a larger convenience sample – so they will hopefully later realize this is why the sample sizes were different.)
- Questions #13 and #14 ask students if they got "similar" results to another group's sample and to the population. If students ask you what "similar" means, you can ask:
  - Do the estimates look close, or very far off? [Let this be open to their interpretation – they can also just say how far off the sample statistic is from the parameter.]

Wrap-up (~ 15-20 min.)

Teacher Instructions Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide. The most important questions/main points (in case you are running out of time) are highlighted.

Large group questions to ask:

- What is the difference between a sample and a population?
- What is the difference between a statistic and a parameter?
- [Project your plot from #6.] Where is this plot centered?
- Do you think that having people name a sample of countries is a biased method of sampling?
  - Why/why not?
  - [*If it turns out that the convenience sample estimates happened to be centered at 71*]: If I had asked you to name 20 countries off the top of your head, how would this plot look different?
- What does it mean for a sampling method to be unbiased?
- Is random sampling an unbiased sampling method?
  - How can you tell based on your plot of results from random sampling (from question #16)?
- What was your answer to question #20? Explain.

After discussion, mention: In real life, we do not have access to the entire population and we usually only take one sample. Rather, we have to be able to trust that our *method* of sampling will tend to produce representative samples of the population and estimates that are unbiased.

IF EXTRA TIME only:

- ☐ What are some examples of polls you have read about in the media that are based on samples?
- ☐ What are some examples of bad sampling you have seen in the media?
- ☐ What are some examples of good sampling you have seen in the media?

Recent election polls may come up as a topic of conversation. Sometimes the polls predict the results correctly, and sometimes they do not. You can ask students to think about why polls are sometimes wrong (e.g., they sample from landlines, older people are more likely to have landlines, etc.)

Appendix D2: Strength Shoe lesson plan

Unit 3, Lesson 2
Strength Shoe

Summary

The Strength Shoe activity looks at the Strength Shoe®, a modified athletic shoe. Its manufacturer claims that this shoe can increase a person's jumping ability. It addresses the research question "How can you design a study to evaluate whether the manufacturer's claim about the Strength Shoe® is legitimate?"

This activity targets the misconception that purposefully assigning groups to balance out known confounding variables is an effective way to assign subjects in an experiment in such a way that causal claims can be made. Students explore a purposeful assignment, balancing out subjects with respect to Sex and Height, but then find there is an unmeasured genetic *X-Factor* that may be affecting jumping ability.

Then, students are guided through the process of randomly assigning subjects, and observe how across many random assignments, differences in confounding variables tend to balance out. The class discusses why we can draw cause-and-effect conclusions based on a randomized experiment.

Learning Goals

This activity has the following goals for students:

- Understand why random assignment is better than purposeful assignment.
- Understand that random assignment tends to balance out confounding variables (both observed and unobserved) between groups.
- Understand why random assignment can enable causal claims.

Reading Preparation

Establishing Causation

TinkerPlots files needed

StrengthShoePurposeful.tp
StrengthShoeRandom-1.tp and StrengthShoeRandom-2.tp

LESSON

Preliminary Discussion (~3-5 min.)

Teacher Instructions: Very brief introduction to the activity's context. To shorten the activity, I removed a question from the activity about anecdotal evidence, but you can ask this question in the preliminary discussion. To lead into the activity, you can discuss how there is a need to design a study to see if the manufacturer's claim is legitimate, rather than relying on anecdotal evidence.

Large group questions to ask:

- Have you ever heard of Strength Shoes?
- If your friend who wears StrengthShoes can jump farther than another friend who wears ordinary training shoes, would you consider this compelling evidence that strength shoes increase jumping ability?
  - Why or why not?

Students work together on the entire activity. (~55 min.)

Teacher Instructions: Ask students to go through the entire activity in their groups.

Suggestions for Potential Issues:

- Questions #1-2 ask about random sampling. If students struggle with these, you can ask:
  - What did you learn about random sampling in the last activity?
  - What kinds of conclusions can you make from studies that use random samples?
- Question #3 asks students to examine a table to see if two groups are balanced with respect to sex. Students might say no because the two groups are not 50% females and 50% males. If so, you can ask:
  - How many females are in each group?
  - How many males are in each group?
  - Are the two groups equal to *each other* with respect to sex?
- Questions #4 and #6 ask if the two groups are "roughly equivalent" with respect to confounding variables. This can be subjective, and it's up to them to decide this. If students ask you what "roughly equivalent" means, you can ask:
  (Answer to Question #4: Strength 67.7; Ordinary 68.3… difference is <1)
  (Answer to Question #6: Strength 83%; Ordinary 17%.... they are very different)
  - Do you think the groups are more or less equal, or very different from each other?

- Questions #10 and #15 ask students to predict what a plot of many random assignments look like. If they struggle with this or ask you how they can know this, you can:
    - Ask them to run their sampler a few more times and see what differences they get.
    - Then, ask them to predict what kind of plot they would get if they ran this sampler 100 more times and plotted these differences.
- Questions #13 and 18 ask students to reason about what the plot being centered near 0 implies about the tendency of random assignment to balance out the variable in the two groups. If they struggle with this, you can ask:
    - What does each dot in the plot represent?
    - What would it mean for a dot to be 0?
    - Why does it make sense that this distribution of differences is centered around or near 0?
- Question #21 is about the ability to make cause-and-effect conclusions from a study with random assignment. Students may still be skeptical because random assignment is not perfectly balancing out the groups in one randomization (especially because the sample sizes are small). If you see this, you can ask:
    - Do you think it's possible to get two groups that are perfectly balanced with respect to all confounding variables in a single random assignment?
    - In the long run, does random assignment tend to balance out the groups with respect to confounding variables?
    - If two groups are relatively balanced with respect to confounding variables, do you think one of these confounding variables is likely to be responsible for the difference in jumping ability?
- Question #23 goes back to generalization. If students ask about this or seem to be confusing "making causal claims" with "generalization," you can ask:
    - If we can make a cause-and-effect conclusion, can we apply this claim to all athletes?
    - Do you think these 12 people are representative of the population of all athletes?

      If they say no because of the sample size, you can ask:
        - If you recruited 100 friends you know, do you think these people would be representative of the population of athletes?

Wrap-up (~ 15-17 min.)

Teacher Instructions

Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide:

Large group questions to ask:

- In this study, what is the treatment variable?
- What is the response variable?
- What does it mean to make a causal claim about the treatment and response variable?
- What is a confounding variable?
- How can confounding variables affect our ability to make causal claims?
- Do you think it's a good idea for humans to purposefully balance out groups with respect to known confounding variables? Why or why not?
- In one single random assignment, do you think it's possible to get two perfectly balanced groups with respect to all confounding variables?
- When you did the random assignments across many trials – why were the plots of the differences centered around 0?
- Does random assignment *tend* to balance out confounding variables?
- Why can we make cause-and-effect conclusions when we have random assignment?

Takeaway points to mention after discussion:

- In real life, we do not perform many random assignments, and we do not have access to "unobserved" confounding variables like the *X*-factor.
- Rather, we just have a single sample of subjects that we randomly assign to treatments one time.
- We have to be able to trust that our method of assignment will tend to balance out potential confounding variables – both the ones we know about and the ones we don't.

IF EXTRA TIME

Teacher Instructions
If you have more time, you can talk about the difference between random assignment and random sampling using these questions. But if you don't have time, don't worry because students will have a whole activity for this.

- What is the difference between random assignment and random sampling?
- Did this study use random sampling?
    - How would this affect our potential conclusions?

Appendix D3: Murderous Nurse lesson plan

Unit 3, Lesson 4
Murderous Nurse

Summary

The *Murderous Nurse* activity looks at when Kristen Gilbert worked as a nurse in the intensive care unit of the Veteran's Administration hospital in Northampton, Massachusetts. It addresses the research question "Were deaths more likely to occur on shifts when Kristen Gilbert was working than on shifts when she was not?" Students go through the process of conducting a randomization test for difference in proportions, with little TinkerPlots™ scaffolding because they have already conducted tests like this before. They also consider what types of inferences can/cannot be made given the design of the study. This is an example of a study where there is no random sampling or random assignment

Learning Goals

This activity has the following goals for students:

- Understand how to use a randomization test for difference in proportions to estimate a p-value and draw a conclusion.
- Understand that when an observed result is more extreme than anything seen in 500 randomized trials, the observed result is extremely unlikely (though not impossible) to happen by chance.
- Understand that generalizations cannot necessarily be made when there is no random sampling.
- Understand that cause-and-effect conclusions cannot necessarily be made when there is no random assignment.
- Understand that even when the study does not include random sampling or random assignment, statistically significant results can still provide evidence of a phenomenon worth investigating further.

Reading Preparation

Scope of Inferences

TinkerPlots files needed

Murderous-Nurse.tp

LESSON

Students work together on the activity (~35 min.)

Teacher Instructions:
Introduce the activity: Tell students we are going back to randomization tests, but now we will carry out a randomization test and consider what the design of the study tells us about the inferences we can make.

Ask students to work through the entire activity in their groups.
Ask students to check their answers to questions #1-6 with a group nearby.
If they are having trouble setting up the TinkerPlots, refer them to the Contagious Yawns activity – the model is set up the same way.
If they are done early – tell students to do an internet search for Kristen Gilbert and see what they can find about her story.

Suggestions for Potential Issues

- Make sure students subtract Gilbert – non-Gilbert in #4.
- Students have only just learned explanatory vs. response and may need help with #5 and #6. If they need guidance, you can point them to the "Establishing Causation" reading and ask questions such as:
    - What variable do we want to predict here? (Death or not)
    - Which variable can help us predict it? (Whether or not Gilbert was working)

Answers to #1-6 for your reference:

- Among all 1641 shifts, the percent of shifts in which a death occurred was 4.5%.
- Among the 257 shifts when Gilbert was working, the percent of shifts in which a death occurred was 15.6%.
- Among the 1384 shifts when Gilbert was not working, the percent of shifts in which a death occurred was 2.4%.
- The difference between the percent of shifts in which a death occurred when Gilbert was wor
- king and the percent of shifts in which a death occurred when Gilbert was not working was 13.2%.
- Explanatory variable: Whether or not Gilbert was working (Gilbert/Non-Gilbert)
- Response variable: Whether or not a death occurred (Death/No Death)

More suggestions for potential issues:

- Students may struggle to find the *p*-value because the difference is off the charts. If so, you can ask:
    - Where is the observed result on the plot?
    - How many trials are beyond the observed result?

- Students may struggle with Question #11: "what does one dot in the plot represent?" If so, you can ask them:
    - What is the null model?
    - What is the sampler doing when it is run each time?
    - What statistic is being collected?
- In question #12 ("where is the plot centered"), students confuse the random assignment in a randomization test with the random assignment in the original data collection. If you observe this, you can ask:
    - What are you modeling in this simulation?
        - Why are we randomly assigning the shifts in this model?
    - How were the shifts in the *original* data divided into groups?
        - Was this random?
    - Point out that these are different questions (first you were asking about the null model, next you were asking about the original data collection.)
- The last two questions (#15-16) are perhaps the most important ones given the emphasis on study design and conclusion. If students struggle with these, you can ask:
    - How were the shifts assigned here to "Gilbert" or "not Gilbert"? What does this imply about potential conclusions?
    - How were these 1641 shifts sampled from the population of ICU shifts? What does this imply about potential conclusions?

Wrap-up (~ 20-25 min.)

Teacher Instructions
Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide:

Large group questions to ask:

- What statistic did you collect?
- What does your plot of 500 randomized trials represent?
- Is the observed difference in percentages statistically significant?
- What does it mean that the observed difference is statistically significant?
- How did you answer the research question?
- How were the shifts sampled?
    - What does this imply about conclusions we can make?
    - What does it mean to generalize?
- How were the shifts assigned?
    - What does this imply about conclusions we can make?
    - What does it mean to make causal claims?
    - (If we can't conclude from the study design that Gilbert caused the deaths, what could be some alternative explanations for this significant difference in percentages?

Teacher Instructions: First, give students about 5-7 minutes to discuss these last 3 questions in small groups. After students have discussed, have them share what they talked about.
[If running out of time, then just pose the questions to the large group].

- If we can't generalize to all shifts, and we cannot necessarily conclude that Gilbert *caused* the deaths, what *can* we say?
- Despite the limitations of this study, do you think that this study would still be valuable in a court of law? Why or why not?
- Would it be advisable to conduct a follow-up study where we randomly assign Kristin Gilbert to shifts in order to strengthen our inferences?

Takeaway points to mention after discussion:

- Studies can still be useful even without random sampling or random assignment.
- Even though we can't make generalizations or causal inferences, we CAN say that this observed difference is unlikely to happen by chance. This "raises our eyebrows." We have evidence that something is going on which warrants further investigation. This information was actually used in the trial, and further evidence pointed to Kristen Gilbert's guilt.
- Experiments are ideal for making causal claims, but not always ethical! If a nurse is suspected of murdering patients it would be unethical to a2q ssign her to shifts.

Appendix D4: Survey Incentives lesson plan

Unit 3, Lesson 4

Survey Incentives

Summary

The Survey Incentives activity introduces students to a situation where both random sampling and random assignment are possible. They play the role of statistical consultants, advising the mayor of a town who wants to design a study to answer the research question: "Will offering a $20 incentive to complete a survey increase response rates for residents of Summerfield?"

Students first advise the mayor on how to sample. They explore 4 variables and compare the distribution of these variables to the population distributions, and also collect a statistic from one of the variables to judge if random sampling is unbiased.

Next, students advise the mayor on how to randomly assign. They are given unequal sample sizes, targeting the misconception that it is impossible to do an experiment with two groups of unequal sample sizes. They choose one potential confounding variable and randomly assign across many trials, observing how random assignment tends to balance out confounding variables.

Lastly, they are asked to compare and contrast random sampling with random assignment and explain how they are different.

Learning Goals

This activity has the following goals for students:

- Understand that the best way to sample from a population is to take a random sample, in order for estimates to be unbiased.
- Understand that the best way to assign groups is to randomly assign, which tends to balance out confounding variables.
- Understand the differences between random sampling and random assignment.
- Understand what inferences random sampling and random assignment allow us to make.

Reading Preparation

None. Students have taken Group Quiz #5 prior to this class period.

TinkerPlots files needed

TownSampling.tp
TownAssignment.tp

LESSON

Students work together on the entire activity. (~55 min.)

Teacher Instructions: Introduce the activity: Students will go through the design of a study, playing the role of "statistical consultant." They will start with sampling and then continue with assignment to groups. Ask students to go through the entire activity together and TURN OFF ANIMATION whenever they collect statistics.

Suggestions for potential issues:

Part 1: Sampling

- Question #3 asks students to advise the mayor on how to take a sample from this list. I anticipate many students might just say "randomly sample 26 names." If you see this happening, you may want to encourage students to go a bit further and describe *how* to do this, for example:
    - You said the mayor can sample randomly, but *how* can the mayor take a random sample from the list?
    - What steps would you advise her to take to obtain her random sample? *Students may come up with a way to use TinkerPlots, and that's OK. If they do this, encourage them to describe how they would set up a sampler to do this.*
- Questions #5-8 have students compare distributions of samples to distributions of the population, asking if the distributions are similar. If students ask what "similar" means, let them know they can decide whether the population and sample distributions look more or less like each other, or very different. You can also ask:
    - Do you expect your sample will have similar characteristics to the population?
    - Why or why not?
- Question #13 asks students to examine a plot of statistics taken from random samples to see if random sampling is unbiased. They should have already done this in the Sampling Countries activity, but if they still struggle with this question, you can ask:
    - What does it mean for a sampling method to be unbiased?
    - If this sampling method is unbiased, where do you expect your plot to be centered?

Part 2: Assignment

- Questions #15 and 16 ask students to pick a potential confounding variable and explain why it might confound the results. If they have trouble choosing, you can ask:
    - Which of the three variables do you think might affect whether people respond or not?
    - How do you think people's [age, income, or hours worked] might influence their willingness to respond?
- Question #17 has students randomly assign an odd number of participants. If they say it's not possible because of the uneven groups, you can ask:

- We may not be able to get an even number in each group, but how can you make the group sample sizes as even as possible?

[If anyone gets hung up over the fact that one participant "moved away" so our sample is no longer representative of the population – you can mention that no study is perfect – there is always the potential for some bias, but we can assume that for the most part the population is the same as it was in the census described above Figure 1.]

- Also in question #17, students may just say "randomly assign ___ to 13 subjects and ___ to 12 subjects" without providing detail. If you see this happening, you may want to encourage students to go a bit further and describe *how* to do this, for example:
  - What detailed steps would you advise the mayor to take in order to carry out this random assignment?

    Again, they may describe how to do this in TinkerPlots and that's OK. If they do this, encourage them to describe how they would set up a sampler to do this.
- Question #18 asks if the two groups are similar with respect to the confounding variable they chose. Again - if students ask what "similar" means, let them know they can decide whether the distributions of the two groups look more or less like each other, or very different. You can also ask:
  - Do you expect the two groups will have similar characteristics?
  - Why or why not?
- Question #21 asks if random assignment is an effective method for balancing out the confounding variable. If students say no or appear to struggle with this, you can ask:
  - Do you expect a single random assignment to perfectly balance out the confounding variable?
  - Does random assignment have the tendency to balance out confounding variables?
- Question #24 gets to the main point of this activity: the difference between random sampling and random assignment. If you see students struggling, or they still think the two study designs are the same thing, you can ask:
  - What kind of conclusion did the mayor want to make in the *Sampling* part of this activity?
  - What was the sampler doing in the *TownSampling* file?
  - What kind of conclusion did the mayor want to make in the *Assignment to Groups* part of this activity?
  - What was the sampler doing in the *TownAssignment* file?

Wrap-up (~ 20 min.)

Teacher Instructions Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide:

Large group questions to ask:

About the sampling method portion:

- What variable did you choose to collect statistics for in question #10?
- Where was your plot in #11 centered?

Try to get some answers from people who chose different variables.

- Why did you expect it to be centered at this value?
- What does it mean for a sampling method to be unbiased?
- Why does an unbiased sampling method allow us to generalize to the population?
- Why is random sampling better than having the mayor drop the surveys into mailboxes on her block?

About the treatment assignment portion:

- In this study, what is the treatment variable?
- What is the response variable?
- What confounding variable did you choose to explore in question #15 and why?
- Where was your plot in #20 centered?

Try to get some answers from people who chose different variables.

- Why does it make sense that your plot was centered around 0?
- What is the purpose of using random assignment in this study?

KEY QUESTIONS: Comparing random assignment and random sampling: If you are running short on time, you can skip earlier questions but make sure you get to these 4 highlighted questions below!

- What is the difference between random assignment and random sampling?
- How is the randomness different in each case?
- Why does random sampling allow us to generalize to the population?
- Why does random assignment allow us to make causal claims?

Some takeaway points to mention:

- In real life, we do not perform many random assignments, or take many random samples – we only have one random sample, and/or one random assignment to groups.
- We need to trust that our sampling method will tend to produce unbiased estimates and is likely to provide us with a representative sample.
- We need to trust that our method of assignment to treatments will tend to balance out potential confounding variables – both the ones we know about and the ones we don't.

Appendix E: Observation Form Checklists

The following four observation forms were used by the two observers (one form for each of the four class activities). The forms each contain a checklist with elements that instructors were given for the lesson plan.

The elements have each been numbered as follows:

- Elements that begin with an "L" (e.g., L1, L2, L3) contain questions or concepts for the instructor to address during large group discussion.
- Elements that begin with an "S" (e.g., S1, S2, S3) contain potential questions or issues that the researcher anticipated could arise during small group activity time.
    - For each of these small group potential issues, suggestions were made for ways in which the instructors could deal with these issues. (For example, for element S1, potential suggestions may be labeled S1A, S1B, etc.)

In addition, for each activity, observers were asked to take notes, focusing on the following general questions to consider:

What do students seem to be getting?
Where do students seem to be struggling?
How is the instructor dealing with student questions?

Appendix E1: Lesson Plan Observation Form for *Sampling Countries*

Unit 3, Lesson 1
Sampling Countries

Summary

The *Sampling Countries* activity allows students to explore and compare different methods of sampling. It addresses the research question "Does the sampling method used impact whether the estimation is unbiased?" Students start with a brief discussion of how they could take samples of students from their class, and which methods might be better than others. Then, they move to the "Sampling Countries" portion of the activity, where the goal is to examine different sampling methods for estimating the average life expectancy. They first take convenience samples of size 20, and then random samples of size 10. The idea is to compare the different sampling methods, and explore how random sampling produces unbiased estimates. The sample sizes for the methods differ so that students can explore how a smaller, random sample is better than a larger, convenience sample because the method of random sampling is unbiased.

Learning Goals

This activity has the following goals for students:

- Understand the difference between biased and unbiased sampling methods
- Understand how human convenience sampling may lead to bias.
- Understand that random sampling produces estimates that are unbiased
- Understand that a smaller random sample is preferable to a larger, biased sample.

Reading Preparation

None. Students have taken the IDEA (Inferences from Design Assessment) as a part of Lab #7 as a pretest, without having had any reading background.

TinkerPlots files needed

SamplingCountries.tp

LESSON

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Introduction			
<2 min.	Briefly introduce the activity, say that we will be talking about methods of sampling from a population. Tell students they have approximately 20-25 minutes to get up to #3, and to stop when they get there. They will be giving you a value to enter into your computer on TinkerPlots.	□ L1. Instructor briefly introduces the activity □ L2. Instructor tells students they have about 20-25 minutes for the first part of this activity. *Potential issues:* □ S1. Students ask instructor what is meant by "representative. □ S1A. Instructor asks something like: "What set of 20 countries do you think might be a good snapshot of the collection of all countries of the world?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Students work on first part of the activity. Instructor will stop after students have given their values.			
~ 25 min	Plot the values in the case table. When the plot is ready, tell students they can move on with the activity and then work through the end.	☐ L3. Instructor plots averages on TinkerPlots for students. ☐ L4. Instructor asks students to continue working on the activity through to the end.	
		Potential Issues ☐ S2. Plot is actually centered at 71, so is unbiased. ☐ If so, instructor should stop class for discussion and lead a discussion asking: ☐ S2A. Where is this plot centered? ☐ S2B. Where do you think this plot of sample averages *would* have been centered if I had asked you to name the first 20 countries you can recall, without asking you to make the sample "representative"? ☐ Why? ☐ S2C. Do you think this sampling method of naming the first 20 countries you can think of would have tended to over-estimate, or under-estimate the average life expectancy?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Students work on random sampling portion of the activity.			
~30 minutes	Have students work through the rest of the activity until the end.	*Potential Issues:* □ S3. Students ask instructor why the random sampling is happening with sample size of 10 instead of 20 like they did earlier. □ S3A. Instructor responds asking students to focus on the *method* of sampling. □ S4. Students ask what "similar" means (when comparing their samples to other samples and to the population) □ S4A. Instructor asks: "Do the estimates look close, or very far off?"	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
WRAP-UP			
~15-20 min	Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide. The most important questions/main points (in case you are running out of time) are highlighted.	Instructor asks wrap-up questions: ☐ L5. What is the difference between a sample and a population? ☐ L6. What is the difference between a statistic and a parameter? ☐ L7. [Project your plot from #6.] Where is this plot centered? ☐ L8. Do you think that having people name a sample of countries is a biased method of sampling? ☐ L8A. Why/why not? ☐ L8B. [*If it turns out that the convenience sample estimates happened to be centered at 71*]: If I had asked you to name 20 countries off the top of your head, how would this plot look different? ☐ L9. What does it mean for a sampling method to be unbiased? ☐ L10. Is random sampling an unbiased sampling method? ☐ L10A. How can you tell based on your plot? ☐ L11. What did you say to question #20? Explain.	

		Instructor mentions take-away points: ☐ L12. In real life, we do not have access to the entire population and we usually only take one sample. ☐ L13. Rather, we have to be able to trust that our *method* of sampling will tend to produce representative samples of the population and estimates that are unbiased. **IF EXTRA TIME** Instructor asks following discussion questions: ☐ L14. What are some examples of polls you have read about in the media that are based on samples? ☐ L15. What are some examples of bad sampling you have seen in the media? ☐ L16. What are some examples of good sampling you have seen in the media? Subjects that could come up during this discussion: ☐ Election polls ☐ Random digit dialing ☐ Bias in sampling methods, such as: ☐ Landline only vs. cell phones ☐ Nonresponse ☐ Other: _____ ☐ _____	

Appendix E2: Lesson Plan Observation Form for *Strength Shoe*

Unit 3, Lesson 2
Strength Shoe

Summary

The Strength Shoe activity looks at the Strength Shoe®, a modified athletic shoe. Its manufacturer claims that this shoe can increase a person's jumping ability. It addresses the research question "How can you design a study to evaluate whether the manufacturer's claim about the Strength Shoe® is legitimate?"

This activity targets the misconception that purposefully assigning groups to balance out known confounding variables is an effective way to assign subjects in an experiment in such a way that causal claims can be made. Students explore a purposeful assignment, balancing out subjects with respect to Sex and Height, but then find there is an unmeasured genetic *X-Factor* that may be affecting jumping ability.

Then, students are guided through the process of randomly assigning subjects, and observe how across many random assignments, differences in confounding variables tend to balance out. The class discusses why we can draw cause-and-effect conclusions based on a randomized experiment.

Learning Goals

This activity has the following goals for students:

- Understand why random assignment is better than purposeful assignment.
- Understand that random assignment tends to balance out confounding variables (both observed and unobserved) between groups.
- Understand why random assignment can enable causal claims.

Reading Preparation

Establishing Causation

TinkerPlots files needed

StrengthShoePurposeful.tp
StrengthShoeRandom-1.tp and StrengthShoeRandom-2.tp

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Introduction			
<2 min.	Very brief introduction to the activity's context. To shorten the activity, I removed a question from the activity about anecdotal evidence, but you can ask this question in the preliminary discussion. To lead into the activity, you can discuss how there is a need to design a study to see if the manufacturer's claim is legitimate, rather than relying on anecdotal evidence.	□ L1. Instructor briefly introduces the activity *Large group questions to ask:* □ L1A. Have you ever heard of Strength Shoes? □ L1B. If your friend who wears StrengthShoes can jump farther than another friend who wears ordinary training shoes, would you consider this compelling evidence that strength shoes increase jumping ability? □ Why or why not?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Students work on the entire activity together			
~ 55 min	Ask students to go through the entire activity in their groups.	☐ L3. Instructor asks students to work on the whole activity in groups.	
		Potential Issues for questions #1-2 (introduction to activity): ☐ S1. Students struggle with question about why random sampling would be preferred. ☐ Instructor asks: ☐ S1A. What did you learn about random sampling in the last activity? ☐ S1B. What kinds of conclusions can you make from studies that use random samples?	
		Potential Issues for question #3 (comparing Sex variable for purposeful assignment): ☐ S2. Students think that "balanced" groups means the groups must be 50/50. ☐ Instructor asks: ☐ S2A. How many females are in each group? ☐ S2B. How many males are in each group? ☐ S2C. Are the two groups equal to *each other* with respect to sex?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
	Teacher Notes: For purposeful assignment: □ (Answer to Question #4: Strength 67.7; Ordinary 68.3… difference is <1) □ (Answer to Question #6: Strength 83%; Ordinary 17%.... they are very different)	*Potential Issues for questions #4-6 (comparing groups with respect to confounding variables):* □ S3. Students have trouble judging whether values are "roughly equivalent" or not. □ Teacher asks: □ S3A. Do you think the groups are more or less equal, or very different from each other?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential Issues for questions #10 and 15 (predicting what a plot of many random assignments will look like): ☐ S4. Students struggle with trying to predict what a plot of many random assignments will look like ☐ Teacher responds by: ☐ S4A. Asking students to run their sampler a few more times and see what differences they get ☐ S4B. Asking them to predict what kind of plot they would get if they ran this sampler 100 more times and plotted these differences.	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential Issues for questions #13 and 18 (reasoning about what the plot being centered at 0 implies about the tendency of random assignment to balance out confounding variables): ☐ S5. Students struggle with reasoning what the plot being centered at 0 implies about random assignment. ☐ Teacher asks: ☐ S5A. What does each dot in the plot represent? ☐ S5B. What would it mean for a dot to be 0? ☐ S5C. Why does it make sense that this distribution of differences is centered around or near 0?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential issues for question #21 (reasoning whether we can make a cause-and-effect conclusion if random assignment was used and a significant difference was found): ☐ S6. Students struggle with answering this question ☐ S7. Students still skeptical about random assignment being able to balance out confounding variables. ☐ Teacher asks: ☐ S7A. Do you think it's possible to get two groups that are perfectly balanced with respect to all confounding variables in a single random assignment? ☐ S7B. In the long run, does random assignment tend to balance out the groups with respect to confounding variables? ☐ S7C. If two groups are relatively balanced with respect to confounding variables, do you think one of these confounding variables is likely to be responsible for the difference in jumping ability?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential issues for question #22: Going back to whether or not we can generalize to a population □ S8. Students struggle with whether or not they can generalize □ S9. Students confuse "generalization" with "causal claims" (or "random sampling" with "random assignment") □ S10. Students do not think they can generalize, only because of the small sample size. □ Teacher asks: □ S10A. If we can make a cause-and-effect conclusion, can we apply this claim to all athletes? □ S10B. Do you think these 12 people are representative of the population of all athletes? □ S10C. If you recruited 100 friends you know, do you think these people would be representative of the population of athletes?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
WRAP-UP ~15-17 min	Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide:	*Instructor asks large group questions:* ☐ L4. In this study, what is the treatment variable? ☐ L5. What is the response variable? ☐ L6. What does it mean to make a causal claim about the treatment and response variable? ☐ L7. What is a confounding variable? ☐ L8. How can confounding variables affect our ability to make causal claims? ☐ L9. Do you think it's a good idea for humans to purposefully balance out groups with respect to known confounding variables? Why or why not? ☐ L10. In one single random assignment, do you think it's possible to get two perfectly balanced groups with respect to all confounding variables? ☐ L11. When you did the random assignments across many trials – why were the plots of the differences centered around 0? ☐ L12. Does random assignment *tend* to balance out confounding variables? ☐ L13. Why can we make cause-and-effect conclusions when we have random assignment?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Instructor mentions these takeaway points to mention after discussion: ☐ L14. In real life, we do not perform many random assignments, and we do not have access to "unobserved" confounding variables like the *X*-factor. ☐ L15. Rather, we just have a single sample of subjects that we randomly assign to treatments one time. ☐ L16. We have to be able to trust that our method of assignment will tend to balance out potential confounding variables– both the ones we know about and the ones we don't.	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
	IF EXTRA TIME If you have more time, you can talk about the difference between random assignment and random sampling using these questions. But if you don't have time, don't worry because students will have a whole activity for this.	*Teacher asks these questions:* ☐ L17. What is the difference between random assignment and random sampling? ☐ L18. Did this study use random sampling? ☐ How would this affect our potential conclusions?	

Appendix E3: Lesson Plan Observation Form for *Murderous Nurse*

Unit 3, Lesson 4
Murderous Nurse

Summary

The *Murderous Nurse* activity looks at when Kristen Gilbert worked as a nurse in the intensive care unit of the Veteran's Administration hospital in Northampton, Massachusetts. It addresses the research question "Were deaths more likely to occur on shifts when Kristen Gilbert was working than on shifts when she was not?" Students go through the process of conducting a randomization test for difference in proportions, with little TinkerPlots™ scaffolding because they have already conducted tests like this before. They also consider what types of inferences can/cannot be made given the design of the study. This is an example of a study where there is no random sampling or random assignment

Learning Goals

This activity has the following goals for students:

- Understand how to use a randomization test for difference in proportions to estimate a p-value and draw a conclusion.
- Understand that when an observed result is more extreme than anything seen in 500 randomized trials, the observed result is extremely unlikely (though not impossible) to happen by chance.
- Understand that generalizations cannot necessarily be made when there is no random sampling.
- Understand that cause-and-effect conclusions cannot necessarily be made when there is no random assignment.
- Understand that even when the study does not include random sampling or random assignment, statistically significant results can still provide evidence of a phenomenon worth investigating further.

Reading Preparation

Scope of Inferences

TinkerPlots files needed

Murderous-Nurse.tp

LESSON

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Introduction to activity			
~3-5 min.	Introduce the activity. Tell students we are going back to randomization tests, but now we will carry out a randomization test and consider what the design of the study tells us about the inferences we can make. Ask students to work through the entire activity in their groups. Ask students to check their answers to questions #1-6 with a group nearby. If they are having trouble setting up the TinkerPlots, refer them to the Contagious Yawns activity – the model is set up the same way. If they are done early – tell students to do an internet search for Kristen Gilbert and see what they can find about her story.	Observation notes: □ L1. Instructor mentions we are going back to randomization tests □ L2. Instructor mentions that now we will consider what the design of the study tells us about the inferences that we can make □ L3. Instructor asks students to work through the activity in their groups. □ L4. Instructor asks students to check their answers to #1-6 with a group nearby. □ L5. Instructor mentions that if students are done early, they can do a search for Kristen Gilbert and find out about her story.	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Students work together on entire activity (~ 45 min.)			
~	Ask students to check their answers to questions #1-6 with a group near them :	Answers to #1-6 (for your reference only - make a note if you see students having issues with any of these questions): o Among all 1641 shifts, the percent of shifts in which a death occurred was 4.5%. o Among the 257 shifts when Gilbert was working, the percent of shifts in which a death occurred was 15.6%. o Among the 1384 shifts when Gilbert was not working, the percent of shifts in which a death occurred was 2.4%. o The difference between the percent of shifts in which a death occurred when Gilbert was working and the percent of shifts in which a death occurred when Gilbert was not working was 13.2%. o Explanatory variable: Whether or not Gilbert was working (Gilbert/Non-Gilbert) o Response variable: Whether or not a death occurred (Death/No Death)	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential issues for questions #5-6: ☐ S1. Students need guidance or are unsure about how to identify explanatory and response variables. Instructor asks: ☐ S1A. What variable do we want to predict here? (Death) ☐ S1B. Which variable can help us predict it? (Whether or not Gilbert was working)	
		Potential issues for question #11: ☐ S2. Students struggle with Question #11: "what does one dot in the plot represent?" ☐ Instructor asks: ☐ S2A. What is the null model? ☐ S2B. What is the sampler doing when it is run each time? ☐ S2C. What statistic is being collected?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential issue for question #12: □ S3. In question #12 ("where is the plot centered"), students confuse the random assignment in a randomization test with the random assignment in the original data collection. □ Instructor asks: □ S3A. What are you modeling in this simulation? □ Why are we randomly assigning the shifts in this model? □ S3B. How were the shifts in the *original* data divided into groups? □ Was this random? □ S3C. Instructor points out that these are different questions (first she was asking about the null model, next she was asking about the original data collection.)	
		Potential issue for question #13: □ S4. Students struggle to find the *p*-value because the difference is off the charts. □ Instructor asks: □ S4A. Where is the observed result on the plot? □ S4B. How many trials are beyond the observed result?	

~25 min		□ S5. Students struggle with questions #15-16 (the ones about study design). □ Instructor asks: □ S5A. How were the shifts assigned here to "Gilbert" or "not Gilbert"? □ What does this imply about potential conclusions? □ S5B. How were these 1641 shifts sampled from the population of ICU shifts? □ What does this imply about potential conclusions?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
WRAP-UP			
~20-25 min	Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide:	Wrap-up questions Part 1: □ L6. What statistic did you collect? □ L7. What does your plot of 500 randomized trials represent? □ L8. Is the observed difference in percentages statistically significant? □ L9. What does it mean that the observed difference is statistically significant? □ L10. How did you answer the research question? □ L11. How were the shifts sampled? □ L11A. What does this imply about conclusions we can make? □ L11B. What does it mean to generalize? □ L12. How were the shifts assigned? □ L12A. What does this imply about conclusions we can make? □ L12B. What does it mean to make causal claims? □ L12C.(If we can't conclude from the study design that Gilbert caused the deaths, what could be some alternative explanations for this significant difference in percentages?	

	First, give students about 5-7 minutes to discuss these last 3 questions in small groups. After students have discussed, have them share what they talked about. *[If running out of time, then just pose the questions to the large group]*.	Wrap-up questions Part 2: - L13. If we can't generalize to all shifts, and we cannot necessarily conclude that Gilbert *caused* the deaths, what *can* we say? - L14. Despite the limitations of this study, do you think that this study would still be valuable in a court of law? Why or why not? - L15. Would it be advisable to conduct a follow-up study where we randomly assign Kristin Gilbert to shifts in order to strengthen our inferences?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Instructor mentions takeaway points: • L16. Studies can still be useful even without random sampling or random assignment. • L17. Even though we can't make generalizations or causal inferences, we CAN say that this observed difference is unlikely to happen by chance. This "raises our eyebrows." We have evidence that something is going on which warrants further investigation. This information was actually used in the trial, and further evidence pointed to Kristen Gilbert's guilt. L18. Experiments are ideal for making causal claims, but not always ethical! If a nurse is suspected of murdering patients it would be unethical to assign her to shifts.	

Appendix E4: Lesson Plan Observation Form for *Survey Incentives*

Note: For this activity, there were many wrap-up questions. The components of the lesson plan appearing in bold represent the essential parts of the lesson plan that instructors were asked to address as key questions. The components of the lesson plan appearing in italics represent the next most important questions, recommended but not required. The components of the lesson plan that were not bolded indicate suggested components or suggestions for potential issues that might arise and how to address them.

Unit 3, Lesson 4
Survey Incentives

Summary

The Survey Incentives activity introduces students to a situation where both random sampling and random assignment are possible. They play the role of statistical consultants, advising the mayor of a town who wants to design a study to answer the research question: "Will offering a $20 incentive to complete a survey increase response rates for residents of Summerfield?"

Students first advise the mayor on how to sample. They explore 4 variables and compare the distribution of these variables to the population distributions, and also collect a statistic from one of the variables to judge if random sampling is unbiased.

Next, students advise the mayor on how to randomly assign. They are given unequal sample sizes, targeting the misconception that it is impossible to do an experiment with two groups of unequal sample sizes. They choose one potential confounding variable and randomly assign across many trials, observing how random assignment tends to balance out confounding variables.

Lastly, they are asked to compare and contrast random sampling with random assignment and explain how they are different.

Learning Goals

This activity has the following goals for students:

- Understand that the best way to sample from a population is to take a random sample, in order for estimates to be unbiased.

- Understand that the best way to assign groups is to randomly assign, which tends to balance out confounding variables.
- Understand the differences between random sampling and random assignment.
- Understand what inferences random sampling and random assignment allow us to make.

Reading Preparation

None. Students have taken Group Quiz #5 prior to this class period.

TinkerPlots files needed

TownSampling.tp
TownAssignment.tp

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Introduction			
<5 min.	Introduce the activity: Students will go through the design of a study, playing the role of "statistical consultant." They will start with sampling and then continue with assignment to groups. Ask students to go through the entire activity together and TURN OFF ANIMATION whenever they collect statistics.	□ L1. Instructor briefly introduces the activity □ L2. Instructor asks students to turn off animation.	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Students work on the entire activity together			
~ 50 min	Ask students to go through the entire activity in their groups.	☐ L3. Instructor asks students to work on the whole activity in groups.	
Part 1: Sampling			
		Potential Issues for question #3 (describing how to take a random sample): ☐ S1. Students just say "randomly sample 26 names." ☐ Instructor asks: ☐ S1A. You said the mayor can sample randomly, but how can the mayor take a random sample from the list? ☐ S2A. What steps would you advise her to take to obtain her random sample? ☐ S2. Students say they will use TinkerPlots to get a random sample. ☐ Instructor asks them to describe how they would set up a sampler to do this.	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential Issues for questions #5-8 (comparing distributions of samples to the distributions of the population variables): ☐ S3. Students ask what "similar" means. ☐ Teacher asks: ☐ S3A. Do the population and sample look more or less like each other, or are they very different from each other? ☐ S3B. Do you expect your sample will have similar characteristics to the population? ☐ Why/why not?	
		Potential Issues for question #13 (asking students to examine a plot of statistics taken from random samples to see if random sampling is unbiased): ☐ S4. Students struggle with question #13 about whether random sampling appears unbiased based on their plot. ☐ Teacher asks: ☐ S4A. What does it mean for a sampling method to be unbiased? ☐ S4B. If this sampling method is unbiased, where do you expect your plot to be centered?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
Part 2: Assignment to Groups			
		Potential Issues for questions #15-16: ☐ S5. Students struggle to pick a confounding variable for question 15 ☐ S6. Students struggle to explain why their confounding variable would affect results ☐ Teacher asks: ☐ S6A. Which of these three variables do you think might affect whether people respond or not? ☐ S6B. How do you think people's [age, income, or hours worked] might influence their willingness to respond?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential Issues for question #17 (asking students how they would randomly assign 25 participants into 2 groups). ☐ S7. Students say random assignment is not possible with an uneven number of people. ☐ Teacher asks: ☐ S7A. We may not be able to get an even number in each group, but how can you make group sample sizes as even as possible? ☐ S8. Students just say "randomly assign the incentive to 13 subjects and the control to 12 subjects (or vice versa) without providing detail. ☐ Teacher asks: ☐ S8A. What detailed steps would you advise the mayor to take in order to carry out this random assignment? ☐ S9. Students say they will use TinkerPlots to get a random assignment. ☐ S9A. Instructor asks them to describe how they would set up a sampler to do this.	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Potential Issues for question #18 (comparing distributions of the control and treatment groups to see if they are similar with respect to the confounding variable students chose): ☐ S10. Students ask what "similar" means. ☐ Teacher asks: ☐ S10A. Do you expect the two groups will have similar characteristics? ☐ Why or why not?	
		Potential issues for question #21 (asking if random assignment is an effective method for balancing out confounding variables): ☐ S11. Students struggle with answering this question ☐ S12. Students say "no" to this question, despite the fact that random assignment was used. ☐ Teacher asks: ☐ S12A. Do you expect a single random assignment to perfectly balance out the confounding variable? ☐ S12B. Does random assignment have the tendency to balance out confounding variables?	

		Potential issues for question #24: Last question, about summarizing the difference between random sampling and random assignment. □ S13. Students struggle with this question □ S14. Students still cannot differentiate between random sampling and random assignment □ Teacher asks: □ S14A. What kind of conclusion did the mayor want to make in the *Sampling* part of this activity? □ S14B. What was the sampler doing in the *TownSampling* file*?* □ S14C. What kind of conclusion did the mayor want to make in the *Assignment to Groups* part of this activity? □ S14D. What was the sampler doing in the *TownAssignment* file*?*	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
WRAP-UP			
~20 min	Lead a large group discussion of the main ideas of the activity, using the following wrap-up questions as a guide **If you are running out of time, be sure to get to the four KEY QUESTIONS at the end of this activity even if you don't have time for the rest.**	***Part 1: Sampling*** *Instructor asks large group questions:* ☐ L4. What variable did you choose to collect statistics for in question #10? ☐ L5. Where was your plot in #11 centered? ☐ Try to get some answers from people who chose different variables. ☐ L6. Why did you expect it to be centered at this value? ☐ L7. What does it mean for a sampling method to be unbiased? ☐ L8. Why does an unbiased sampling method allow us to generalize to the population? ☐ L9. Why is random sampling better than having the mayor drop the surveys into mailboxes on her block?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Part 2: Assignment *Instructor asks large group questions:* ☐ L10. In this study, what is the treatment variable? ☐ L11. What is the response variable? ☐ L12. What confounding variable did you choose to explore in question #15 and why? ☐ L13. Where was your plot in #20 centered? ☐ Try to get some answers from people who chose different variables. ☐ L14. Why does it make sense that your plot was centered around 0? ☐ L15. What is the purpose of using random assignment in this study?	
		KEY QUESTIONS: ☐ L16. What is the difference between random assignment and random sampling? ☐ L17. How is the randomness different in each case? ☐ L18. Why does random sampling allow us to generalize to the population? ☐ L19. Why does random assignment allow us to make causal claims?	

~TIME	TEACHER INSTRUCTIONS	OBSERVATION CHECKLIST	OBSERVATION NOTES
		Instructor mentions these takeaway points after discussion, if time. (They should have already been mentioned in wrap-ups prior to this.) □ In real life, we do not perform many random assignments, or take many random samples – we only have one random sample, and/or one random assignment to groups. □ We need to trust that our sampling method will tend to produce unbiased estimates and is likely to provide us with a representative sample. □ We need to trust that our method of assignment to treatments will tend to balance out potential confounding variables – both the ones we know about and the ones we don't.	

Appendix F: Group Quiz and Rubric

Appendix F1: Group Quiz

Group Quiz #5

Each student in your group needs to take the role of writer/recorder for a portion of the exam (as indicated). S/he will be responsible for helping the group come to consensus and also for writing the group's agreed upon response.
Use for 1 - 2

In January 2016, researchers conducted the Gallup-Healthways Well-Being Index survey with a random sample of 347,915 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia[7]. The survey asked adults about many health habits and well-being factors, such as alcohol consumption. Among the survey's findings were that moderate drinkers (1-14 alcoholic drinks per week) were significantly less likely to have had a depression diagnosis and more likely to experience positive emotions than non-drinkers (0 drinks per week).

Writer/Recorder A: _____

1. The headline of the article reads: "In U.S., Moderate Drinkers Have Edge in Emotional Health." (In other words, the headline claims that in the United States, those who drink moderately tend to have better emotional health than those who do not drink.) Given the study design, is this an appropriate headline? Explain.

[7] Nekvasil, N. & Liu, D. (2016). *Gallup.* "In U.S., moderate drinkers have edge in emotional health." Retrieved from: http://www.gallup.com/poll/188816/moderate-drinkers-edge-emotional-health.aspx?g_source=CATEGORY_WELLBEING&g_medium=topic&g_campaign=tiles

2. Suppose you encounter a media article from an online news outlet reporting the results of this study. The article recommends that American adults consider drinking alcohol in moderate amounts in order to lower their levels of depression and increase positive emotions. Given how the study was designed, is this an appropriate recommendation to make? Explain.

Use for 3 - 4

Does the size of the bowl affect how much ice cream you eat? Since it is known that people tend to eat most of what they serve themselves, obesity researchers were interested in examining whether the size of a bowl unknowingly affects how much ice cream a person serves him/herself.

Their study consisted of 42 nutrition experts in Massachusetts who attended an ice cream social. The participants were randomly assigned either a smaller (17 oz) or a larger (34 oz) bowl, and then each participant self-served the amount of ice cream in her/his bowl. After serving themselves, each nutritionist's bowl was weighed and the amount of ice cream was recorded (in ounces). The response variable of interest is the amount of ice cream in the smaller and larger bowls. A randomization test revealed that participants who had the larger bowl ate significantly more ice cream, on average, than participants who had the smaller bowl ($p < .05$).

Writer/Recorder B: _____

3. Based on the design of this study, is it likely that factors other than bowl size may explain the difference between the average amount of ice cream in the larger and smaller bowl groups? Explain.

4. The results from this study showed that those who had the larger bowl ate significantly more ice cream than those with the smaller bowl. Is this result generalizable to all nutritionists in Massachusetts? Explain.

Use for 5 - 6

A reporter from an online news outlet hires you as a statistical consultant. She wants to make sure that the headlines she is publishing for her articles are accurate and reflect appropriate conclusions. She is currently writing an article about the following study:

Educational policy experts accessed records of all students who applied to medical school at public universities in the United States in 2014. A random sample of 250 student records was collected and analyzed, looking at admission status and undergraduate grade point average (GPA). Two groups of students were compared: those who were offered admission to medical school and those who were denied admission. The average undergraduate GPA was compared between groups. A significant difference in averages was found ($p < 0.05$), with higher average GPA for students who were offered admission.

Writer/Recorder C: _____

5. The reporter proposes the headline: "New study: Higher grades get you into medical school at public universities in the U.S." Based on the design of this study, would you recommend that she publish this headline? Explain.

6. The reporter also has another choice of headline: "Admission to public medical schools in the United States associated with higher college grades." Based on the design of this study, would you recommend that she publish this headline? Explain.

Appendix F2: Group Quiz Rubric

Group Quiz #5: RUBRIC

Each student in your group needs to take the role of writer/recorder for a portion of the exam (as indicated). S/he will be responsible for helping the group come to consensus and also for writing the group's agreed upon response.
Use for 1 - 2

In January 2016, researchers conducted the Gallup-Healthways Well-Being Index survey with a random sample of 347,915 adults, aged 18 and older, living in all 50 U.S. states and the District of Columbia[8]. The survey asked adults about many health habits and well-being factors, such as alcohol consumption. Among the survey's findings were that moderate drinkers (1-14 alcoholic drinks per week) were significantly less likely to have had a depression diagnosis and more likely to experience positive emotions than non-drinkers (0 drinks per week).

Writer/Recorder A: _____

7. The headline of the article reads: "In U.S., Moderate Drinkers Have Edge in Emotional Health." (In other words, the headline claims that in the United States, those who drink moderately tend to have better emotional health than those who do not drink.) Given the study design, is this an appropriate headline? Explain.

   Yes – this is making a claim generalizing the association found in this study to the U.S. population. This claim is appropriate given that a random sample was used in the study.

   To get the full point the student must say "yes" and provide a reasonable explanation referring to the random sampling, such as:
   - *Random sampling was used in the study, allowing us to generalize to the US adult population*
   - *The sample is representative of the US population, as it was taken randomly; therefore we can claim that this association applies to the US population.*

---

[8] Nekvasil, N. & Liu, D. (2016). *Gallup.* "In U.S., moderate drinkers have edge in emotional health." Retrieved from: http://www.gallup.com/poll/188816/moderate-drinkers-edge-emotional-health.aspx?g_source=CATEGORY_WELLBEING&g_medium=topic&g_campaign=tiles

Ways to get half credit (0.5 points):

- Students read the headline and mistakenly think it is making a causal claim, which leads them to say no. For example:

- *"The headline claims that drinking will give you an edge in emotional health, and we cannot say this because adults were not randomly assigned to groups."*

- Students correctly reason that random sampling is the study design that is needed, but do not realize that random sampling was used. For example:

- *"No, we cannot make this claim about the general U.S. population, because the sample was not randomly selected from the US population."*

- Students recognize that the sample was taken from all 50 states and therefore claim it's representative, but do not make specific reference to the random sampling. For example:

- *"Yes, we can make this claim generalizing to the US population, because it was a large sample taken from all 50 states."*

- Students say it is OK to make this claim of association because the study is observational/random assignment was not used, but fail to recognize the headline is trying to make a generalization. For example:

- *"Yes, we can make this claim because even though random assignment was not used, the headline is only making a claim of association between drinking and emotional health."*

Do NOT give any credit if all they talk about is the sample size. Example:

- *Yes, we can make this claim generalizing to the U.S. population because the sample size was 347,915.*

8. Suppose you encounter a media article from an online news outlet reporting the results of this study. The article recommends that American adults consider drinking alcohol in moderate amounts in order to lower their levels of depression and increase positive emotions. Given how the study was designed, is this an appropriate recommendation to make? Explain.

   No – although moderate drinkers have better emotional health than non-drinkers, there could be confounding variables. This study did not use random assignment, so we cannot make causal claims about the effect of drinking on emotional health.

   To get the full point the student must say "no" and provide a reasonable explanation such as:

   - *This was an observational study*
   - *No random assignment was used/is possible*

- *There could be confounding variables that explain these results*

Ways to get half credit (0.5 points):

- *Students correctly recognize that this recommendation requires a causal claim we cannot make, but do NOT reference the lack of random assignment or potential for confounding. For example:*
    - *"No, we cannot necessarily claim that drinking moderately will lead to better mental health because of the study design."*
    - *"No, we cannot make cause-and-effect statements like this based on the study design."*
- *Students correctly reason that just because a significant difference was found in a group, that does not mean we can guarantee that the result will be the same for each individual person. But they do <u>not</u> reference the observational nature of the study or lack of random assignment. For example:*
    - *"No, just because moderate drinkers have less depression on average than non-drinkers, does not mean that drinking moderately will help every person to prevent depression."*
    - *"No – although we can say that in the US population, moderate drinkers are less likely to develop depression than non-drinkers, this does not mean that for any one person, drinking moderately will lower depression risk.*

-NO credit if "yes", such as:

- "Yes, because random sampling was used so we can conclude that moderate drinking leads to better emotional health" (i.e. confusing random sampling with random assignment.)

<u>Use for 3 - 4</u>

Does the size of the bowl affect how much ice cream you eat? Since it is known that people tend to eat most of what they serve themselves, obesity researchers were interested in examining whether the size of a bowl unknowingly affects how much ice cream a person serves him/herself.

Their study consisted of 42 nutrition experts in Massachusetts who attended an ice cream social. The participants were randomly assigned either a smaller (17 oz) or a larger (34 oz) bowl, and then each participant self-served the amount of ice cream in her/his bowl. After serving themselves, each nutritionist's bowl was weighed and the amount of ice cream was recorded (in ounces). The response variable of interest is the amount of ice cream in the smaller and larger bowls. A randomization test revealed that participants who had the larger

bowl ate significantly more ice cream, on average, than participants who had the smaller bowl ($p < .05$).

Writer/Recorder B: _____

9. Based on the design of this study, is it likely that factors other than bowl size may explain the difference between the average amount of ice cream in the larger and smaller bowl groups? Explain.

   No, because the random assignment balances out other variables (i.e. confounding variables) that could explain this difference.

   To get the full point they should say "no" and make reference to the fact that random assignment was used, with a reasonable explanation such as:
   - *Random assignment should balance out factors other than bowl size that could explain the difference.*
   - *Because of the experimental design using random assignment, confounding variables should not likely be a concern.*
   - *The random assignment balanced out the groups so that they are similar in terms of other factors that could explain difference in bowl size (e.g., appetite, age, weight, diet)*

   Ways to get half credit (0.5 points):

   - *Students recognize that random assignment was used, but still do not recognize that the random assignment should balance out the confounding factors. For example:*
     - *"Yes, other factors could explain the difference. Even though random assignment was used, there still could be factors such as appetite and diet that affect people's serving size."*
     - *"Yes, there can still be differences in confounding variables between groups even after the random assignment."*

10. The results from this study showed that those who had the larger bowl ate significantly more ice cream than those with the smaller bowl. Do you think that this result is generalizable to all nutritionists in Massachusetts? Explain.

No, because the participants did not consist of a random sample. They consisted of nutritionists at an ice cream social, and this sample might not be representative of all nutritionists in Massachusetts.

To get the full point, students should say "no", and give a reasonable explanation such as:

- *there was no random sampling of nutritionists in Massachusetts*
- *the nutritionists in the sample were not representative of the population*
- *there is bias in the sampling method, because the nutritionists were attending an ice cream social*

For this question, I cannot think of ways to earn half credit, but if you encounter an answer that is on the right track but not quite getting there, please ask me.

Do NOT give any credit if the answer ONLY refers to the sample size but makes no reference to the sampling method. For example:

- "No, because there were only 42 nutritionists in the sample so we cannot generalize to all nutritionists in Massachusetts." (Without talking about how it was an ice cream social, or referring to bias/lack of random sampling.)

NO credit for"yes", such as:

- "Yes, because random assignment was used, so we can make generalizations…" (i.e. confusing random sampling with random assignment.)

<u>*Use for 5 - 6*</u>

A reporter from an online news outlet hires you as a statistical consultant. She wants to make sure that the headlines she is publishing for her articles are accurate and reflect appropriate conclusions. She is currently writing an article about the following study:

Educational policy experts accessed records of all students who applied to medical school at public universities in the United States in 2014. A random sample of 250 student records was collected and analyzed, looking at admission status and undergraduate grade point average (GPA). Two groups of students were compared: those who were offered admission to medical school and those who were denied admission. The average undergraduate GPA was compared between groups. A significant difference in averages was found ($p < 0.05$), with higher average GPA for students who were offered admission.

Writer/Recorder C: _____

11. The reporter proposes the headline: “New study: Higher grades get you into medical school at public universities in the U.S.” Based on the design of this study, would you recommend that she publish this headline? Explain.

No – although higher GPAs are associated with getting into medical school, we cannot make causal claims because this is an observational study and GPA cannot be randomly assigned. Other variables such as student motivation could explain why students with higher GPA are more likely to get into medical school.

To get the full point, students should say “no” and provide a reasonable explanation that references the lack of random assignment or potential for confounding, such as:

- *no causal claims can be made because this is an observational study*
- *no causal claims can be made because random assignment was not used here*
- *other confounding variables (such as motivation, study habits, etc. could explain why students with higher GPAs are more likely to get into medical school.*

Ways to get half credit (0.5 points):

- *Students say we cannot make causal claims from this study, so the headline is wrong, but do NOT reference the lack of random assignment or potential for confounding. For example:*
  - *“No, we cannot necessarily conclude that higher GPA causes people to be more likely to get into medical school.”*
  - *“No, we cannot make cause-and-effect statements like this based on the study design.”*
- *Students interpret this claim as making a generalization only, and miss the fact that it’s making a causal statement. But they still reason correctly about the random sampling allowing for this generalization. For example:*
  - *“Yes, a random sample was taken from student records, so we can make this claim generalizing to US public medical schools.*
  - *“Yes, we can claim that those with higher GPAs are more likely to get into US public medical schools, because a random sample of records from this population was taken.”*

12. The reporter also has another choice of headline: “Admission to public medical schools in the United States associated with higher college grades.” Based on the design of this study, would you recommend that she publish this headline? Explain.

- *Yes. The records were a random sample of all students who applied to medical school at public universities in the United States in 2014. This should provide a representative sample, so we can generalize to the population as the headline claims.*
- *To get the full point, students should say "yes" with a reasonable explanation such as:*
    - *the records were a random sample, so they should be representative of the population of US public medical school applications*
    - *the headline is making a generalization, which we can do because the records were sampled randomly from the population of interest*
- *Ways to get half credit (0.5 points):*
- *Students mistakenly think the headline is making a causal claim, but correctly explain that this is not possible because of the lack of random assignment. For example:*
    - *No, we cannot claim that getting higher grades will help admission into US medical schools, because this was an observational study/no random assignment was used.*
    - *No, we cannot claim that admission to public medical schools is caused by higher grades, because there could be other factors/confounding variables such as student major, experience, etc.*
- *Students say it is OK to make this claim of association because the study is observational/random assignment was not used, but fail to recognize the headline is trying to make a generalization. For example:*
    - *Yes, we can publish this headline because it is not making a causal claim, just one of association. Random assignment was not used, but this headline is OK because it's not trying to make causal claims.*
    - *This was an observational study, so we can only make claims about association, not causation. So this headline is OK.*

Appendix G: Lab Assignment and Rubric

Appendix G1: Lab Assignment

Lab Assignment 08

Part 1: IDEA-B (Inferences from Design Assessment)

Please go to the following website to take this multiple choice assessment (22 total questions) online. You have seen these questions before, and we would like to see how you are reasoning about concepts of study design and conclusions after going through the Unit 3 activities so far. Please answer each question to the best of your knowledge and ability.

http://z.umn.edu/3264lab8part1

Part 2: Peanut Allergies

In this lab assignment you will be presented with excerpts from two separate studies of peanut allergies[9]. The researchers who conducted these studies used different study designs. You will be asked to read excerpts of these studies and consider the inferences and conclusions that can be made based on the study design.

Remember there are two primary questions that you should ask when evaluating a study's design:(1) How were the study participants selected from the population?; and (2) How were the selected study participants assigned to conditions?

---

[9] Sicherer, S. H., Wood, R. A., Stablein, D., Lindblad, R., Burks, A. W., Liu, A. H., Jones, S. M., Fleischer, D. M., Leung, D. Y., & Sampson, H. A. (2010). Maternal consumption of peanut during pregnancy is associated with peanut sensitization in atopic infants. *Journal of Allergy and Clinical Immunology, 126*(6), 1191–1197.

Slomski, A. (2015). Consuming—Not Avoiding—Peanuts Leads to Fewer Peanut Allergies in Kids. *Journal of the American Medical Association, 313*(16), 1609–1609.

Excerpt #1

Consider the following excerpt from a study reported in the *Journal of the American Medical Association*.

Consuming—Not Avoiding—Peanuts Leads to Fewer Peanut Allergies in Kids

High-risk children who consumed peanut products from infancy until they were 5 years old were significantly less likely to develop a peanut allergy than those who avoided peanuts, according to the LEAP randomized trial.

The 640 infants in the trial were recruited to be in the study based on the following criteria: they were aged 4 to 11 months at enrollment, and all had severe eczema, egg allergy, or both. Participants in each cohort were randomly assigned to consume a peanut protein–containing bar or to avoid peanuts.

Among the 530 infants in (one) cohort, the prevalence of peanut allergy at 60 months was 13.7% in the avoidance group and 1.9% in the consumption group. The absolute difference in risk of 11.8% represents an 86.1% relative reduction in the prevalence of peanut allergy.

1. Identify the explanatory variable in this study.

2. Identify the response variable in this study.

3. The excerpt indicates that children who consumed peanut products from infancy until they were 5 years old were “significantly less likely to develop a peanut allergy than those who avoided peanuts”. Explain what the term “significantly” means in this context.

4. What is the population of interest in this study?

5. Based on the study design, does it appear that the researchers can generalize findings to this population? Explain.

6. The title of the article assumes a causal relationship between the treatment and response variables. Given the study design, is such a claim appropriate? Explain.

Excerpt #2

Consider the following excerpt from the *Journal of Allergy and Clinical Immunology*.

> **Maternal consumption of peanut during pregnancy is associated with peanut sensitization in atopic infants**
>
> To identify factors associated with peanut sensitization.... we evaluated 503 infants 3 to 15 months of age (mean, 9.4 months). These infants were recruited based on having no previous diagnosis of peanut allergy.
>
> Multivariate analysis including clinical, laboratory, and demographic variables showed frequent peanut consumption during pregnancy ($p < .001$),... male sex ($p = .02$), and nonwhite race ($p = .02$) to be the primary factors associated with peanut (allergy).

7. In this study, several groups were compared. Identify all the explanatory variables given in the excerpt. (Hint: there are three).

8. Identify the response variable in the study.

9. What is the population of interest in this study?

10. Based on the study design, does it appear that the researchers can generalize findings to this population? Explain.

11. For each of the three explanatory variables you identified in question #7, were the participants assigned to groups?

12. What does your answer to question #11 imply about the types of inferences researchers can or cannot make based on the study results?

13. A classmate tells you that if the 503 infants in this study had been randomly sampled from the population, we could determine whether frequent peanut consumption during pregnancy causes a higher incidence of allergies. Is your classmate's statement correct? Explain.

14. A colleague of yours is pregnant and says that based on the results described in excerpt #2, she definitely wants to avoid eating peanuts during pregnancy so that her child will have a smaller chance of developing peanut sensitivity. Based on the design of the study described in excerpt #2, what would you tell her?

Appendix G2: Lab Rubric

Lab Assignment 08

RUBRIC

Part 2: Peanut Allergies

In this lab assignment you will be presented with excerpts from two separate studies of peanut allergies[10]. The researchers who conducted these studies used different study designs. You will be asked to read excerpts of these studies and consider the inferences and conclusions that can be made based on the study design.

Remember there are two primary questions that you should ask when evaluating a study's design:(1) How were the study participants selected from the population?; and (2) How were the selected study participants assigned to conditions?

Model answers are below. The most important ideas that students should understand in this lab are:

- Recognizing that in order to generalize to an appropriate population of interest, random sampling from that population is needed.
- Recognizing that in order to make causal claims, random assignment to groups is needed.
- Ability to discern from the text describing a study whether random sampling, random assignment, or neither was used.

---

[10] Sicherer, S. H., Wood, R. A., Stablein, D., Lindblad, R., Burks, A. W., Liu, A. H., Jones, S. M., Fleischer, D. M., Leung, D. Y., & Sampson, H. A. (2010). Maternal consumption of peanut during pregnancy is associated with peanut sensitization in atopic infants. *Journal of Allergy and Clinical Immunology, 126*(6), 1191–1197.

Slomski, A. (2015). Consuming—Not Avoiding—Peanuts Leads to Fewer Peanut Allergies in Kids. *Journal of the American Medical Association, 313*(16), 1609–1609.

- Ability to distinguish between issues of generalization (random sampling necessary) and issues of cause-and-effect (random assignment necessary).

Notes about extra scaffolding questions:

- *Questions 1, 2, 7, and 8 (asking about explanatory/response variables) are there as scaffolding to help students later identify whether the explanatory variable was randomly assigned, and whether one can make a causal claim about the explanatory and response variables.*
- *Questions 4 and 9(asking about identifying the population) are there as scaffolding to help students later identify whether the sample was taken randomly from the population of interest.*
- *Question 3 (about what it means to have significance) is there to help students review what they have learned in Unit 2 about significance, and it is also scaffolding for question #6 that asks them about whether causal claims can be made. (You cannot claim causation without a significant association to begin with.)*

Holistic scoring:

(3) Answers exhibit a **complete understanding** of the concepts in the assignment. There are no errors in student's statistical reasoning. The responses are clear and correct.

(2) Answers exhibit a **near complete understanding** of the assignment. There are perhaps minor errors in student's statistical reasoning or the responses are slightly unclear or incorrect.

(1) Answers exhibit **some understanding** of the assignment. There are errors in student's statistical reasoning or the responses are unclear or incorrect.

(0) Answers exhibit **little to no understanding** of the assignment. There are fundamental errors in student's statistical reasoning or the responses are unclear or incorrect.

Excerpt #1

Consider the following excerpt from a study reported in the *Journal of the American Medical Association*.

> **Consuming—Not Avoiding—Peanuts Leads to Fewer Peanut Allergies in Kids**
>
> High-risk children who consumed peanut products from infancy until they were 5 years old were significantly less likely to develop a peanut allergy than those who avoided peanuts, according to the LEAP randomized trial.
>
> The 640 infants in the trial were recruited to be in the study based on the following criteria: they were aged 4 to 11 months at enrollment, and all had severe eczema, egg allergy, or both. Participants in each cohort were randomly assigned to consume a peanut protein–containing bar or to avoid peanuts.
>
> Among the 530 infants in (one) cohort, the prevalence of peanut allergy at 60 months was 13.7% in the avoidance group and 1.9% in the consumption group. The absolute difference in risk of 11.8% represents an 86.1% relative reduction in the prevalence of peanut allergy.

1. Identify the explanatory variable in this study.

   Peanut consumption (consuming a peanut protein-containing bar or avoiding peanuts.)

2. Identify the response variable in this study.

   Peanut allergy (or whether or not the infants had peanut allergy)

3. The excerpt indicates that children who consumed peanut products from infancy until they were 5 years old were "significantly less likely to develop a peanut allergy than those who avoided peanuts". Explain what the term "significantly" means in this context.

   "Significantly" means that the difference observed in peanut allergy prevalence was unlikely to happen by chance (p-value likely smaller than .05).

4. What is the population of interest in this study?

   Infants aged 4-11 months with severe eczema, egg allergy, or both. (Also OK to just say infants ages 4-11 months, as this could arguably be the population of interest.)

5. Based on the study design, does it appear that the researchers can generalize findings to this population? Explain.

   No – the participants were recruited to be in the study, so they were likely not a random sample. They might not be representative of the population of interest.

6. The title of the article assumes a causal relationship between the treatment and response variables. Given the study design, is such a claim appropriate? Explain.

   Yes – the participants were randomly assigned to consume or avoid peanuts, so potential confounding variables should be balanced out, allowing us to make causal claims.

Excerpt #2

Consider the following excerpt from the *Journal of Allergy and Clinical Immunology*.

> **Maternal consumption of peanut during pregnancy is associated with peanut sensitization in atopic infants**
>
> To identify factors associated with peanut sensitization.... we evaluated 503 infants 3 to 15 months of age (mean, 9.4 months). These infants were recruited based on having no previous diagnosis of peanut allergy.
>
> Multivariate analysis including clinical, laboratory, and demographic variables showed frequent peanut consumption during pregnancy ($p < .001$),... male sex ($p = .02$), and nonwhite race ($p = .02$) to be the primary factors associated with peanut (allergy).

7. In this study, several groups were compared. Identify all the explanatory variables given in the excerpt. (Hint: there are three).

   Peanut consumption during pregnancy, sex, and race.

8. Identify the response variable in the study.

   Peanut allergy

9. What is the population of interest in this study?

   Infants ages 3-15 months without previous diagnosis of peanut allergy (or also OK to say infants ages 3-15 months as this could have arguably been the population of interest.)

10. Based on the study design, does it appear that the researchers can generalize findings to this population? Explain.

    No – infants were recruited for this study. They might not be representative of the population.

11. For each of the three explanatory variables you identified in question #7, were the participants assigned to groups?

    No – peanut consumption during pregnancy, sex, and race were all observed by the researchers, but not controlled.

12. What does your answer to question #11 imply about the types of inferences researchers can or cannot make based on the study results?

    We cannot make causal claims because random assignment here was not done, so there could be confounding variables that explain these relationships.

13. A classmate tells you that if the 503 infants in this study had been randomly sampled from the population, we could determine whether frequent peanut consumption during pregnancy causes a higher incidence of allergies. Is your classmate's statement correct? Explain.

    No – random sampling has to do with generalization, not with making causal claims. Random assignment is what is needed for making causal claims.

14. A colleague of yours is pregnant and says that based on the results described in excerpt #2, she definitely wants to avoid eating peanuts during pregnancy so that her child will have a smaller chance of developing peanut sensitivity. Based on the design of the study described in excerpt #2, what would you tell her?

    Several potential issues could be discussed here:
    - *The pregnant women were not randomly assigned to eat peanuts or not, so we cannot make causal claims and assume that peanut avoidance will lead to a smaller chance of peanut sensitivity.*

- *It is unclear whether the findings would apply to the colleague, as the sample was not randomly selected it is not clear to what population you can generalize.*

Appendix H: Correspondence with reviewers of the IDEA assessment and blueprint

Appendix H1: Initial invitation e-mail to reviewers

Subject:

Invitation to be an expert reviewer for my dissertation research – Univ. of Minnesota

Dear ____,

I am a doctoral student beginning my dissertation research project which focuses on students' understanding of study design and conclusions. I am in the Statistics Education graduate program at the University of Minnesota, where I am working with my advisers, Dr. Bob delMas and Dr. Andy Zieffler. In particular, I am developing learning activities to help students distinguish between random sampling and random assignment and the role that these study designs play in the scope of inferences that can be made. One of the tools I am using to evaluate student learning outcomes from these activities is an assessment that consists mostly of items modified from previously existing assessments that have been used in statistics education research (e.g., GOALS, ARTIST, CAOS).

Because of your expertise in the area of statistics education, I am writing to request your assistance in this project which would involve reviewing a 17-item forced choice assessment. If you agree to participate in my research, I would ask you to indicate the extent to which you believe each item aligns with its intended learning goal, and to give any suggestions you have for modifying the items. This will be done in a Microsoft Word document and should take around 25-30 minutes.

If you agree to participate as an expert reviewer, I will send you the 17 items and the instructions for reviewing them no later than *February 15*th. I would like to receive feedback by *February 29th* (this will give you 2 weeks). Please feel free to ask me any questions that you have. I sincerely hope that you will be able to contribute to my research.

Please let me know whether or not you are able to participate.

Thank you,

Elizabeth Fry
PhD Candidate in Quantitative Methods in Education
University of Minnesota

Appendix H2: E-mail of instructions for reviewers after each agreed to participate

Subject:

Re: Invitation to be an expert reviewer for my dissertation research – Univ. of Minnesota

Dear ____,

Thank you for agreeing to review the assessment for my dissertation research. The goal of this assessment will be to evaluate introductory statistics students' understanding of study design and conclusions. Specifically, students should understand how random sampling allows for unbiased estimation and generalization, and how random assignment helps to balance out confounding variables which allows for causal claims to be made. I am attaching two documents: a blueprint of the assessment goals and the 22-item assessment. (I had previously said 17 items were on it, but based on feedback from my advisor, several of these items were turned into short item sets instead.)

Please provide feedback on the attached 22-item assessment document by using the Microsoft Word "comments" feature, and feel free to mark up suggested changes using "track changes." As you provide feedback, please keep in mind the following questions:

- What suggestions do you have for improving an item, keeping in mind the item's intended assessment goal?
- What suggestions do you have for improving clarity and wording of the items and responses?
- Do you think there are any important assessment goals missing? Do any of the assessment goals seem redundant?

In order to give me enough time to modify the assessment and post it online for students to take later this spring semester, I would like to receive your feedback by *February 29* if possible.

I truly appreciate your assistance in my research. Please let me know if you have any questions.

Thank you,

Elizabeth Fry

Doctoral Candidate

Appendix I: IDEA blueprint

Unbiased Estimation: items #1-9
1-2: (Two-item set): Ability to identify the sample and the population to which inferences can be made.
3: Ability to understand what it means to make an appropriate generalization to a population, using sample data.
4: Ability to understand the factors that allow (or do not allow) a sample of data to be representative of the population.
5: Ability to understand when sample estimates may be biased due to lack of a representative sample.
6: Ability to understand that a small random sample is preferable to a larger, biased sample.
7: Ability to understand that random sampling is preferable to non-random methods of sampling for a sample to be representative of the population.
8: Ability to understand that sample statistics vary from sample to sample.
9: Ability to recognize that random sampling is the most salient issue when using a sample to generalize to a population.

Establishing Causation: items 10-22
10: Ability to determine what type of study was conducted (observational or experimental).
11: Ability to understand that a randomized experiment is needed to answer research questions about causation.
12-15 (Four-item set): Ability to distinguish between statements that make causal claims and statements that make association-only claims
16: Ability to understand that correlation does not imply causation.
17: Ability to understand how a confounding variable may explain the association between an explanatory and response variable
18: Ability to understand the purpose of random assignment in an experiment: To make groups comparable with respect to all other confounding variables.
19-21 (three-item set): Ability to understand that random assignment is the best way to balance out groups with respect to confounding variables.
22: Ability to recognize when a randomized experiment is the most salient research design for a particular research question.

Appendix J: IDEA instrument with tables of responses

Students who completed each IDEA version (pretest or posttest)

	Section				
	1	2	3	4 (online)	Total
IDEA-A (pretest)	39	32	24	36	131
IDEA-B (posttest)	39	30	28	33	130

Use for questions 1 and 2: The Pew Research Center surveyed a nationally representative group of 1,002 American adults in 2013. Of these adults, 21% have had an email or social networking account compromised.

1. Identify the sample used in this study.

a. The sample is all American adults in 2013.
b. The sample is the 21% of American adults that have had an email or social networking account compromised.
c. The sample is the 1,002 American adults surveyed.

	Answer option			
Section	a	b	c	Condition
1	2.6	7.7	**89.7**	A (N = 39)
	5.1	0.0	**94.9**	B (N = 39)
2	0.0	18.8	**81.3**	A (N = 32)
	0.0	10.0	**90.0**	B (N = 30)
3	0.0	4.2	**95.8**	A (N = 24)
	3.6	3.6	**92.9**	B (N = 28)
4	5.6	2.8	**91.7**	A (N = 36)
	6.1	3.0	**90.9**	B (N = 33)
Overall	2.3	8.4	**89.3**	A (N = 131)
	3.8	3.8	**92.3**	B (N = 130)

2. Identify the population about which the Pew Research Center can make inferences based on the survey results.

a. The population is all American adults in 2013.
b. The population is the 21% of American adults that have had an email or social networking account compromised.
c. The population is the 1,002 American adults surveyed.

	Answer option			
Section	a	b	c	Condition
1	**48.7**	25.6	25.6	A (*N* = 39)
	61.5	15.4	23.1	B (*N* = 39)
2	**40.6**	21.9	37.5	A (*N* = 32)
	66.7	13.3	20.0	B (*N* = 30)
3	**37.5**	45.8	16.7	A (*N* = 24)
	67.9	14.3	17.9	B (*N* = 28)
4	**33.3**	38.9	27.8	A (*N* = 36)
	69.7	12.1	18.2	B (*N* = 33)
Overall	**40.5**	32.1	27.5	A (*N* = 131)
	66.2	13.8	20.0	B (*N* = 130)

3. Administrators at Central High School randomly sampled 100 students from the student body, and found that the high school students who had studied a foreign language tended to score significantly higher, on average, on the SAT than the high school students who had not studied a foreign language. Which of the following statements correctly represents a *generalization* that can be made to an *appropriate population* of interest?

a. High school students who study a foreign language have significantly higher average SAT scores than those who do not study a foreign language.
b. Students at Central High School who study a foreign language have significantly higher average SAT scores than those who do not study a foreign language.
c. Out of the 100 sampled students from Central High School, those who study a foreign language have significantly higher average SAT scores than those who do not study a foreign language.

	Answer option			
Section	a	**b**	c	Condition
1	28.2	**20.5**	51.3	A (N = 39)
	5.1	**74.4**	20.5	B (N = 39)
2	25.0	**28.1**	46.9	A (N = 32)
	16.7	**56.7**	26.7	B (N = 30)
3	20.8	**33.3**	45.8	A (N = 24)
	14.3	**64.3**	21.4	B (N = 28)
4	27.8	**16.7**	55.6	A (N = 36)
	9.1	**51.5**	39.4	B (N = 33)
Overall	26.0	**23.7**	50.4	A (N = 131)
	10.8	**62.3**	26.9	B (N = 130)

4. A college official conducted a survey of students currently living in dormitories to learn about their preference for single rooms, double rooms, or multiple (more than two people) rooms in the dormitories on campus. Out of 5,000 total students who live in dormitories on campus, a random sample of 500 first-year students was selected and the official received survey results from 160 of these students.

Which of the following does **NOT** affect the college official's ability to generalize the survey results to all dormitory students at this college?

a. Only 500 students were sent the survey.

b. The survey was sent to only first-year students.

c. Of the 500 students who were sent the survey, only 160 responded.

d. All of the above present a problem for generalizing the results to all dormitory students at this college.

	Answer option				
Section	**a**	b	c	d	Condition
1	**10.3**	12.8	0.0	76.9	A (N = 39)
	35.9	12.8	7.7	43.6	B (N = 39)
2	**9.4**	3.1	3.1	84.4	A (N = 32)
	23.3	16.7	23.3	36.7	B (N = 30)
3	**12.5**	12.5	16.7	58.3	A (N = 24)
	42.9	14.3	3.6	39.3	B (N = 28)
4	**0.0**	13.9	11.1	75.0	A (N = 36)
	33.3	3.0	9.1	54.5	B (N = 33)
Overall	**7.6**	10.7	6.9	74.8	A (N = 131)
	33.8	11.5	10.8	43.8	B (N = 130)

5. A local television station in a city with a population of 500,000 recently conducted a poll where they invited viewers to call in and voice their support or opposition to a controversial referendum that was to be voted on in an upcoming election. Over 10,000 people responded, with 67% opposed to the referendum. The TV station announced that they are convinced that the referendum will be defeated in the election.

Select the answer below that indicates whether the TV station's announcement is valid or invalid, and why.

a. Valid, because the sample size is large enough to represent the population.
b. Valid, because 67% is far enough above 50% to predict a majority vote.
c. Invalid, because the sample is too small given the size of the population.
d. Invalid, because the sample may not be representative of the population.

	Answer option				
Section	a	b	c	**d**	Condition
1	2.6	5.1	5.1	**87.2**	A (*N* = 39)
	0.0	2.6	7.7	**89.7**	B (*N* = 39)
2	12.5	6.3	25.0	**56.3**	A (*N* = 32)
	3.3	6.7	0.0	**90.0**	B (*N* = 30)
3	0.0	0.0	16.7	**83.3**	A (*N* = 24)
	10.7	0.0	0.0	**89.3**	B (*N* = 28)
4	13.9	5.6	22.2	**58.3**	A (*N* = 36)
	15.2	6.1	3.0	**75.8**	B (*N* = 33)
Overall	7.6	4.6	16.8	**71.0**	A (*N* = 131)
	6.9	3.8	3.1	**86.2**	B (*N* = 130)

6. Two surveys were conducted to determine the percentage of higher education institutions in Texas that have a recycling program for waste. Survey A sent postcards to the deans of all 208 higher education institutions in Texas. Half (104) of the deans sent them back, and 91% of those that returned the postcards said that their institution recycled. Survey B used a random sample of 50 higher education institutions and contacted the deans of each college by phone. Out of the 50 deans, 20 of them (40%) said their institution recycled. Select the response below that indicates which survey is most likely to provide an unbiased estimate of the proportion of all higher education institutions in Texas that recycle and why.

a. Survey A, because the sample size is larger.
b. Survey A, because all of the deans were contacted.
c. Survey B, because the deans were contacted by phone rather than mail.
d. Survey B, because the sample was randomly selected.

	Answer option				
Section	a	b	c	**d**	Condition
1	7.7	23.1	23.1	**46.2**	A (N = 39)
	0.0	7.7	17.9	**74.4**	B (N = 39)
2	21.9	9.4	15.6	**53.1**	A (N = 32)
	0.0	3.3	6.7	**90.0**	B (N = 30)
3	12.5	4.2	33.3	**50.0**	A (N = 24)
	3.6	0.0	0.0	**96.4**	B (N = 28)
4	22.2	22.2	16.7	**38.9**	A (N = 36)
	3.0	6.1	6.1	**84.8**	B (N = 33)
Overall	16.0	16.0	21.4	**46.6**	A (N = 131)
	1.5	4.6	8.5	**85.4**	B (N = 130)

7. The science club at a large middle school has 25 members. The members want to survey a sample of 125 students to estimate the percentage of students at the school who plan to submit a science-fair project. Each member of the club asks 5 friends the following question: "Will you be submitting a project to the science fair this year?" Of all the friends, 76% replied with a "yes." Which of the following is a reason why the sample selection described is biased?

a. A sample of friends is not likely to be representative of students at the school.

b. The club members did not survey every student at the school.

c. A sample of 125 students is too small to be representative of students at the school.

d. The percentage of students who plan to submit a project is not equal to 50%.

	Answer option				
Section	**a**	b	c	d	Condition
1	**97.4**	0.0	2.6	0.0	A (N = 39)
	97.4	0.0	2.6	0.0	B (N = 39)
2	**87.5**	3.1	9.4	0.0	A (N = 32)
	96.7	0.0	0.0	3.3	B (N = 30)
3	**91.7**	0.0	4.2	4.2	A (N = 24)
	100.0	0.0	0.0	0.0	B (N = 28)
4	**80.6**	11.1	8.3	0.0	A (N = 36)
	93.9	3.0	3.0	0.0	B (N = 33)
Overall	**89.3**	3.8	6.1	0.8	A (N = 131)
	96.9	0.8	1.5	0.8	B (N = 130)

8. In a study, Researcher A took a random sample of 25 college students and found the mean number of times they went out to eat during the last week was 4.1. In another study, Researcher B took a random sample of 25 students from the same college and found the mean number of times they went out to eat during the last week was 3.7. What is the best explanation for why the samples taken by Researcher A and Researcher B did not produce the same mean?

a. The sample means varied because they are calculated from small samples.

b. The sample means varied because the samples were not representative of all college students.

c. The sample means varied because each sample is a different subset of the population.

	Answer option			
Section	a	b	**c**	Condition
1	25.6	5.1	**69.2**	A (*N* = 39)
	28.2	7.7	**64.1**	B (*N* = 39)
2	9.4	12.5	**78.1**	A (*N* = 32)
	20.0	0.0	**80.0**	B (*N* = 30)
3	37.5	8.3	**54.2**	A (*N* = 24)
	17.9	7.1	**75.0**	B (*N* = 28)
4	38.9	5.6	**55.6**	A (*N* = 36)
	18.2	6.1	**75.8**	B (*N* = 33)
Overall	27.5	7.6	**64.9**	A (*N* = 131)
	21.5	5.4	**73.1**	B (*N* = 130)

9. The dean of a college would like to determine the feelings of students concerning a new registration fee that would be used to upgrade the recreational facilities on campus. To collect data, the dean hires graduate students to stand outside the library and ask everyone who walks by the entrance to fill out a survey. Results show that of the 100 students who filled out the survey, students who live on campus are significantly more opposed to the fee than those who live off campus. Later, the student newspaper prints the headline: "Across the college, students who live on campus are more opposed to new registration fee than off-campus students." What is the biggest problem with printing this headline?

a) The sample size was too small.
b) This was an observational study.
c) Random assignment was not used in the study.
d) Random sampling was not used in the study.

	Answer option				
Section	a	b	c	**d**	Condition
1	25.6	7.7	17.9	**48.7**	A (*N* = 39)
	10.3	7.7	15.4	**66.7**	B (*N* = 39)
2	28.1	21.9	6.3	**43.8**	A (*N* = 32)
	6.7	3.3	30.0	**60.0**	B (*N* = 30)
3	29.2	0.0	8.3	**62.5**	A (*N* = 24)
	0.0	10.7	28.6	**60.7**	B (*N* = 28)
4	36.1	11.1	2.8	**50.0**	A (*N* = 36)
	6.1	3.0	18.2	**72.7**	B (*N* = 33)
Overall	29.8	10.7	9.2	**50.4**	A (*N* = 131)
	6.2	6.2	22.3	**65.4**	B (*N* = 130)

10. Suppose a researcher wanted to determine if aspirin reduces the chance of a heart attack. The researcher studied 500 patients who visited a regional hospital in the last year. Half (250) of the patients were randomly assigned to take aspirin every day and the other half to take a placebo everyday. Then after a certain length of time, the percentage of heart attacks for the patients who took aspirin every day and the percentage for those who did not take aspirin every day were reported. What type of study did the researcher conduct?

a. Observational
b. Experimental
c. Survey

	Answer option			
Section	a	**b**	c	Condition
1	0.0	**100.0**	0.0	A (*N* = 39)
	7.7	**92.3**	0.0	B (*N* = 39)
2	0.0	**96.9**	3.1	A (*N* = 32)
	3.3	**96.7**	0.0	B (*N* = 30)
3	8.3	**91.7**	0.0	A (*N* = 24)
	14.3	**82.1**	3.6	B (*N* = 28)
4	8.3	**88.9**	2.8	A (*N* = 36)
	12.1	**87.9**	0.0	B (*N* = 33)
Overall	3.8	**94.7**	1.5	A (*N* = 131)
	9.2	**90.0**	0.8	B (*N* = 130)

11. A researcher is studying the relationship between a vitamin supplement and cholesterol level. Which of the following would allow the researchers to establish that taking a vitamin supplement regularly causes a change in cholesterol level?

a. Measure the cholesterol levels of 100 patients and record whether or not they regularly take the vitamin supplement.
b. Randomly assign 50 patients to regularly take a vitamin supplement, 50 to take a placebo, and compare their cholesterol levels.
c. Send a survey that asks about vitamin supplements and cholesterol levels to a random sample of 100 patients.

	Answer option			
Section	a	**b**	c	Condition
1	2.6	**97.4**	0.0	A (*N* = 39)
	2.6	**97.4**	0.0	B (*N* = 39)
2	3.1	**96.9**	0.0	A (*N* = 32)
	0.0	**96.7**	3.3	B (*N* = 30)
3	8.3	**91.7**	0.0	A (*N* = 24)
	3.6	**96.4**	0.0	B (*N* = 28)
4	8.3	**91.7**	0.0	A (*N* = 36)
	3.0	**93.9**	3.0	B (*N* = 33)
Overall	5.3	**94.7**	0.0	A (*N* = 131)
	2.3	**96.2**	1.5	B (*N* = 130)

Use the following for items 12-15: For each of the following media article headlines, determine whether the statement on the right indicates a claim of association only, or a claim of causation.

12. a. **Association** b. Causation Number of Facebook friends linked to size of brain regions.

	Answer option		
Section	**a**	b	Condition
1	**97.4**	2.6	A (N = 39)
	100.0	0.0	B (N = 39)
2	**90.6**	9.4	A (N = 32)
	93.3	6.7	B (N = 30)
3	**91.7**	8.3	A (N = 24)
	96.4	3.6	B (N = 28)
4	**91.7**	8.3	A (N = 36)
	93.9	6.1	B (N = 33)
Overall	**93.1**	6.9	A (N = 131)
	96.2	3.8	B (N = 130)

13. a. Association b. **Causation** Daily exercise improves mental performance

	Answer option		
Section	a	**b**	Condition
1	5.1	**94.9**	A (N = 39)
	5.1	**94.9**	B (N = 39)
2	9.4	**90.6**	A (N = 32)
	10.0	**90.0**	B (N = 30)
3	20.8	**79.2**	A (N = 24)
	10.7	**89.3**	B (N = 28)
4	5.6	**94.4**	A (N = 36)
	6.1	**93.9**	B (N = 33)
Overall	9.2	**90.8**	A (N = 131)
	7.7	**92.3**	B (N = 130)

14. a. Association b. **Causation** Cell phone radiation leads to death in honeybees.

	Answer option		
Section	a	**b**	Condition
1	2.6	**97.4**	A (*N* = 39)
	10.3	**89.7**	B (*N* = 39)
2	6.3	**93.8**	A (*N* = 32)
	16.7	**83.3**	B (*N* = 30)
3	25.0	**75.0**	A (*N* = 24)
	14.3	**85.7**	B (*N* = 28)
4	13.9	**86.1**	A (*N* = 36)
	15.2	**84.8**	B (*N* = 33)
Overall	10.7	**89.3**	A (*N* = 131)
	13.8	**86.2**	B (*N* = 130)

15. a. **Association** b. Causation Cat owners tend to be more educated than dog owners.

	Answer option		
Section	**a**	b	Condition
1	**100.0**	0.0	A (*N* = 39)
	100.0	0.0	B (*N* = 39)
2	**93.8**	6.3	A (*N* = 32)
	90.0	10.0	B (*N* = 30)
3	**95.8**	4.2	A (*N* = 24)
	100.0	0.0	B (*N* = 28)
4	**88.9**	11.1	A (*N* = 36)
	97.0	3.0	B (*N* = 33)
Overall	**94.7**	5.3	A (*N* = 131)
	96.9	3.1	B (*N* = 130)

16. Researchers conducted a survey of 1,000 randomly selected adults in the United States and found a strong, positive, statistically significant correlation between income and the number of containers the adults reported recycling in a typical week.

Can the researchers conclude that higher income causes more recycling among U.S. adults? Select the best answer from the following options.

a. No, the sample size is too small to allow causation to be inferred.
b. No, the lack of random assignment does not allow causation to be inferred.
c. Yes, the statistically significant result allows causation to be inferred.
d. Yes, the sample was randomly selected, so causation can be inferred.

	Answer option				
Section	a	**b**	c	d	Condition
1	25.6	**30.8**	7.7	35.9	A (N = 39)
	5.1	**87.2**	0.0	7.7	B (N = 39)
2	40.6	**18.8**	12.5	28.1	A (N = 32)
	0.0	**76.7**	6.7	16.7	B (N = 30)
3	29.2	**45.8**	12.5	12.5	A (N = 24)
	3.6	**85.7**	10.7	0.0	B (N = 28)
4	44.4	**16.7**	13.9	25.0	A (N = 36)
	18.2	**57.6**	3.0	21.2	B (N = 33)
Overall	35.1	**26.7**	11.5	26.7	A (N = 131)
	6.9	**76.9**	4.6	11.5	B (N = 130)

17. A research team wanted to study the relationship between completing an internship during college and students' future earning potential. From the same graduating class, they selected a random sample of 80 students who completed an internship and 100 students who did not complete an internship and examined their salaries 5 years after graduation. They found a significantly higher mean salary for the internship group than for the non-internship group. Which of the following is a reasonable statement based on this study?

a. More students should take internships because having an internship produces a higher salary.
b. Another variable, such as student major, could explain the difference in mean salaries.
c. You cannot draw any valid conclusions because the samples are not the same size.

	Answer option			
Section	a	**b**	c	Condition
1	20.5	**66.7**	12.8	A (*N* = 39)
	15.4	**84.6**	0.0	B (*N* = 39)
2	4.2	**83.3**	12.5	A (*N* = 32)
	13.3	**70.0**	16.7	B (*N* = 30)
3	4.2	**83.3**	12.5	A (*N* = 24)
	7.1	**92.9**	0.0	B (*N* = 28)
4	25.0	**55.6**	19.4	A (*N* = 36)
	6.1	**75.8**	18.2	B (*N* = 33)
Overall	17.6	**67.2**	15.3	A (*N* = 131)
	10.8	**80.8**	8.5	B (*N* = 130)

18. A research study randomly assigned participants into two groups. One group was given Vitamin E to take daily. The other group received only a placebo pill. The research study followed the participants for eight years. After the eight years, the proportion of each group that developed a particular type of cancer was compared.

What is the primary reason that the study used random assignment?

a. To ensure that the groups are likely to be similar in all respects except for the level of Vitamin E.

b. To ensure that a person is not likely to know whether or not they are getting the placebo.

c. To ensure that the study participants are likely to be representative of the larger population.

	Answer option			
Section	**a**	b	c	Condition
1	**41.0**	20.5	38.5	A (*N* = 39)
	84.6	7.7	7.7	B (*N* = 39)
2	**25.0**	31.3	43.8	A (*N* = 32)
	73.3	6.7	20.0	B (*N* = 30)
3	**37.5**	33.3	29.2	A (*N* = 24)
	78.6	10.7	10.7	B (*N* = 28)
4	**25.0**	30.6	44.4	A (*N* = 36)
	66.7	12.1	21.2	B (*N* = 33)
Overall	**32.1**	28.2	39.7	A (*N* = 131)
	76.2	9.2	14.6	B (*N* = 130)

Use for questions 19-21: An instructor is going to conduct an experiment in his statistics class to compare the effect of 4 different exam preparation methods on student understanding. There are 40 students in the class. Indicate whether (Yes) or not (No) each of the following methods for distributing the students to the 4 exam preparation methods will allow the instructor to balance out groups with respect to potential confounding variables, so that the instructor can attribute any differences in average scores between the groups to the effect of the exam preparation methods.

19. a. Yes b. **No** Ask students to sit in four different groups of 10, then randomly assign each group to an exam preparation method (for example, group 1 is randomly assigned method 3, group 2 is randomly assigned method 1, group 3 is randomly assigned method 4 and group 4 is randomly assigned method 2).

	Answer option		
Section	a	**b**	Condition
1	38.5	**61.5**	A (N = 39)
	26.3	**73.7**	B (N = 38)
2	40.0	**60.0**	A (N = 30)
	26.7	**73.3**	B (N = 30)
3	37.5	**62.5**	A (N = 24)
	35.7	**64.3**	B (N = 28)
4	27.8	**72.2**	A (N = 36)
	15.2	**84.8**	B (N = 33)
Overall	35.7	**64.3**	A (N = 129)
	25.6	**74.4**	B (N = 129)

20. a. **Yes** b. No Assign a unique number from 1 to 40 to each student, then using a random sequence of the numbers 1 to 40, assign the students with the first 10 numbers in the sequence to the first exam preparation method, the students with the second set of 10 numbers to the second exam preparation method, and so on.

	Answer option		
Section	**a**	b	Condition
1	**89.7**	10.3	A (N = 39)
	92.1	7.9	B (N = 38)
2	**93.8**	6.3	A (N = 32)
	93.3	6.7	B (N = 30)
3	**83.3**	16.7	A (N = 24)
	85.7	14.3	B (N = 28)
4	**83.3**	16.7	A (N = 36)
	93.9	6.1	B (N = 33)
Overall	**87.8**	12.2	A (N = 131)
	91.5	8.5	B (N = 129)

21. a. Yes b. **No** Assign the exam preparation method as students walk into class, giving the first exam preparation method to the first 10 students and the second exam preparation method to the next 10 students, and so on.

	Answer option		
Section	a	**b**	Condition
1	33.3	**66.7**	A (N = 39)
	23.7	**76.3**	B (N = 38)
2	30.0	**70.0**	A (N = 30)
	23.3	**76.7**	B (N = 30)
3	50.0	**50.0**	A (N = 24)
	21.4	**78.6**	B (N = 28)
4	44.4	**55.6**	A (N = 36)
	15.2	**84.8**	B (N = 33)
Overall	38.8	**61.2**	A (N = 129)
	20.9	**79.1**	B (N = 129)

22. Conducting an experiment with random assignment to treatments is most appropriate for answering which of the following questions?

a. Do students learn more if they listen to music while studying?
b. How has the population of the United States changed in the last 100 years?
c. What is the average height of 20 children in a kindergarten class?
d. What percentage of high school students in California eat breakfast before going to school?

	Answer option				
Section	**a**	b	c	d	Condition
1	**79.5**	2.6	2.6	15.4	A (N = 39)
	89.5	2.6	0.0	7.9	B (N = 38)
2	**87.5**	3.1	3.1	6.3	A (N = 32)
	90.0	0.0	3.3	6.7	B (N = 30)
3	**79.2**	0.0	8.3	12.5	A (N = 24)
	96.4	0.0	0.0	3.6	B (N = 28)
4	**63.9**	5.6	11.1	19.4	A (N = 36)
	93.9	0.0	6.1	0.0	B (N = 33)
Overall	**77.1**	3.1	6.1	13.7	A (N = 131)
	92.2	0.8	2.3	4.7	B (N = 129)

Appendix K: Frequency and Percent of Students with Item Response Patterns for IDEA items

Sampling Items

			Item response pattern[a]				Mc Nemar's exact test p-value
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post	
1	(Two-item set): Ability to identify the sample and the population to which inferences can be made.	125	2 (1.6%)	8 (6.4%)	10 (8.0%)	105 (84.0%)	0.814
2		125	32 (25.6%)	11 (8.8%)	42 (33.6%)	40 (32.0%)	<.001
3	Ability to understand what it means to make an appropriate generalization to a population, using sample data.	125	38 (30.4%)	8 (6.4%)	58 (46.4%)	21 (16.8%)	<.001
4	Ability to understand the factors that allow (or do not allow) a sample of data to be representative of the population.	125	76 (60.8%)	9 (7.2%)	39 (31.2%)	1 (0.8%)	<.001
5	Ability to understand when sample estimates may be biased due to lack of a representative sample.	125	11 (8.8%)	6 (4.8%)	26 (20.8%)	82 (65.6%)	<.001
6	Ability to understand that a small random sample is preferable to a larger, biased sample.	125	13 (10.4%)	5 (4.0%)	54 (43.2%)	53 (42.4%)	<.001
7	Ability to understand that random sampling is preferable to non-random methods of sampling for a sample to be representative of the population.	125	3 (2.4%)	1 (0.8%)	11 (8.8%)	110 (88.0%)	0.006

			Item response pattern[a]				Mc Nemar's exact test p-value
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post	
8	Ability to understand that sample statistics vary from sample to sample.	125	21 (16.8%)	12 (9.6%)	22 (17.6%)	70 (56.0%)	0.121
9	Ability to recognize that random sampling is the most salient issue when using a sample to generalize to a population.	125	30 (24.0%)	14 (11.2%)	31 (24.8%)	50 (40.0%)	0.016

[a]Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest

Assignment Items

			Item response pattern[a]				McNemar's exact test p-value
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post	
10	Ability to determine what type of study was conducted (observational or experimental).	125	2 (1.6%)	10 (8.0%)	5 (4.0%)	108 (86.4%)	0.302
11	Ability to understand that a randomized experiment is needed to answer research questions about causation.	125	3 (2.4%)	1 (0.8%)	4 (3.2%)	117 (93.6%)	0.375
12	(Four-item set): Ability to distinguish between statements that make causal claims and statements that make association-only claims	125	1 (0.8%)	4 (3.2%)	8 (6.4%)	112 (89.6%)	0.388
13		125	3 (2.4%)	7 (5.6%)	9 (7.2%)	106 (84.8%)	0.804
14		125	5 (4.0%)	12 (9.6%)	9 (7.2%)	99 (79.2%)	0.664
15		125	0 (0.0%)	4 (3.2%)	7 (5.6%)	114 (91.2%)	0.549
16	Ability to understand that correlation does not imply causation.	125	21 (16.8%)	7 (5.6%)	69 (55.2%)	28 (22.4%)	<.001
17	Ability to understand how a confounding variable may explain the association between an explanatory and response variable	125	15 (12.0%)	10 (8.0%)	26 (20.8%)	74 (59.2%)	0.011

			Item response pattern[a]				McNemar's exact test p-value
Item	Measured Learning Outcome	*n*	Incorrect	Decrease	Increase	Pre & Post	
18	Ability to understand the purpose of random assignment in an experiment: To make groups comparable with respect to all other confounding variables.	125	21 (16.8%)	7 (5.6%)	64 (51.2%)	33 (26.4%)	<.001
19	(three-item set): Ability to understand that random assignment is the best way to balance out groups with respect to confounding variables.	122	15 (12.3%)	16 (13.1%)	27 (22.1%)	64 (52.5%)	0.126
20		124	0 (0.0%)	10 (8.1%)	14 (11.3%)	100 (80.6%)	0.541
21		122	15 (12.3%)	10 (8.1%)	33 (27.0%)	64 (52.5%)	<.001
22	Ability to recognize when a randomized experiment is the most salient research design for a particular research question.	124	7 (5.6%)	3 (2.4%)	18 (14.5%)	96 (77.4%)	0.001

[a]Incorrect = incorrect on both the pretest and posttest; Decrease = correct pretest, incorrect posttest; Increase = incorrect pretest, correct posttest; Pre & Post = correct on both the pretest and posttest

Appendix L: Qualitative codebook

The lab assignment and group quizzes were coded according to the following behaviors. The behaviors are split into three categories: (1) Misconceptions/incorrect thinking, (2) Correct Thinking, and (3) Ambiguity. In addition, extra behaviors having to do with specific questions were coded, as described below.

Incorrect thinking/Misconceptions **(I)**

Misunderstandings about which study designs help with which types of conclusions **(TC)**

- **I-TC-RSC Bringing up only random sampling/lack thereof when the question is about causation (e.g., saying you can make causal claims because a random sample was taken)**
    - Examples:
        - "*When there is no random sampling from the explanatory variables and the response variables we cannot conclude our results that one thing caused another…*"
        - "*No, this claim [of causation] is not appropriate because the study does not utilize random sampling*.
- **I-TC-RAG Bringing up only random assignment/lack thereof when the question is about generalization (e.g., saying you can generalize to a population because random assignment was used.)**
    - Examples:
        - "*Because there was no random assignment, the results of this study cannot be generalized to the population as a whole*."
        - "*You cannot use these findings to generalize to this population because they were not randomly assigned*."
        - *"The experiment can only make a causation claim depends if the experiment is random sampling.*"
- **I-TC-BOTHG Saying you need both random sampling AND random assignment to generalize**
    - Examples:
        - *"Based on the study design, there is not random assignment, nor random sampling. Therefore, the researchers cannot generalize the findings to the population*."
        - "*No. It does not appear that researchers can generalize their findings to this population, as there was no random sampling or assignment present within the study.*"

- **I-TC-BOTHC Saying you need both random sampling AND random assignment to make causal claims**
    - Example: "*This [causal] claim would only be appropriate if random sampling occurred along with random assignment.*"
- **I-TC-CLAIM Confusing the meaning of "generalize" with the meaning of "causal claims"**
    - Example: *"Based on the study design, the researchers may generalize the findings to this population because the study was carried out in random assignment and confounding variables might have been balanced out. This means that the real cause of the study (peanut consumption) can be concluded*."
- **I-TC-NOCC Not believing causal claims can be made even though random assignment was used** (still saying confounding variables can affect results, despite acknowledging the fact that random assignment was used)
    - Example: "*Yes, there are other confounding variables that come into play which may have played a part in the results. For example, perhaps ice cream lovers randomly got assigned the 34 oz. bowl but regardless of the fact, they would have scooped more ice cream anyways."*

Incorrect beliefs about sample size **(SS)**

- **I-SS-UNEVEN Saying that unequal sample sizes in two groups do not allow for any conclusions**
    - (Not observed in assessment answers, but documented in class activity observations.)
- **I-SS-LARGEN Saying we can generalize due to the large sample size**
    - Examples:
        - "*I would say yes [you can generalize] because out of the 630 infants in the sample, a large majority shown significant results in the model, of course a larger sample size would help you draw conclusions about the general population better though.*"
        - "*Yes [we can generalize], because of the significant difference and the relatively large sample size.*"
- **I- SS-SMALLN Saying we can't generalize (or make any conclusion) only because of small sample size**
    - Example: "*Although there was a large difference between the two groups, especially in the second cohort of 98 infants, the number that were involved in this study might be a little too small to say that the results can transfer to a larger population.*"

Difficulty understanding study descriptions (SD)

- **I-SD-RECRS Difficulty recognizing from study description whether random sampling was used:** (e.g., assuming the sample was random when in fact it wasn't.)
    - Example: "*Based on the study it does appear that researchers can generalize findings to this population. This is because the article states that the LEAP trial was randomized, meaning that random sampling from within the population occurred so we can generalize findings to the population of high risk children." [Describing a study that used random assignment, but not random sampling.]*
- **I-SD-RECRA Difficulty recognizing from study description whether random assignment was used** (e.g., assuming random assignment was done, when it was not; or assuming random assignment was not done, when it was.)
    - Example: "*No, the study doesn't state that it was a random assignment. Therefore there could be unbalanced unknown confounding variables." [Describing a study where random assignment was done.]*
    - Example: *"Yes, there are other factors [that] could explain the amount of ice cream in each bowl. One variable that could affect the study is how hungry each participant is.." [Describing a study where random assignment was done, not acknowledging that the random assignment was done.]*

Examples of correct reasoning (C)
Understanding that random sampling helps to make generalizations, or that generalizations cannot be made if the sample is not representative of the population (SG)

- **C-SG-RSGEN Pointing out that random sampling is relevant for generalizing to a wider population**
    - Examples:
        - "*Yes, [you can publish this headline] because this is a generalization to the population that was randomly sampled.*"
        - "*No, this result is not generalizable to all nutritionists in Massachusetts because there was no random sampling.*"
- **C-SG-SCHAR Mentioning that the sample can have characteristics that make it different from the population**
    - Examples:
        - "*You cannot use these findings to generalize to a wider population because the study recruited the study participants based on certain conditions, which included infants with no prior diagnosis of peanut allergies instead of using random sampling.*"
        - *"It does not appear that the researchers would be able to make generalizations about their findings to the whole infant population because the infants in the study were chosen specifically because they*

were already high risk (had eczema and/or egg allergy). These findings would not be fair to generalize about babies without eczema and egg allergies."

Understanding that random assignment helps to make causal claims, or that causal claims cannot be made if confounding variables could explain differences between groups (AC)

- **C-AC-RACC Pointing out that random assignment is relevant for making causal claims**
    - Examples:
        - *"Yes, based on the study design a causal inference is appropriate because the participants were randomly assigned to either one of the two groups."*
        - *"It doesn't specify that they were randomly assigned; Therefore the researchers can't make a causal claim about the study."*
- **C-AC-CONFV Mentioning that confounding variables can make two groups different from each other**
    - Examples:
        - "*For example, say there is a gene in some individuals of a given race that makes them less susceptible to peanut allergies. Then it would not matter how much peanuts that mother from that given race chose to consume. And so because the confounding variables are almost endless, and random assignment was not present, this study cannot claim any form of causations.*"
        - "*No, there are many confounding factors not taken into consideration. There may be factors associated with moderate drinkers that influence their emotional well-being other than drinking moderately.*"

Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (WHY)

- **C-WHY-RS Explaining *why* random sampling helps us to generalize**
    - Examples:
        - "*These finding are likely not generalizable, the study didn't specify that they used a random sample of infants. For example, parents may have been more likely to participate in a study if a peanut allergy runs in their family, which would be a confounding factor that could skew the results."*

        - *"Generalization can be used because they took a random sample from the population for this study. A random sample is usually a good representation of the overall population."*
- **C-WHY-RA Explaining *why* random assignment helps us to make causal claims**
    - Examples:
        - "*Random assignment assures that all other confounding variables have been balanced out, and thus, the only identifiable difference between infants was the variable being manipulated (that being whether or not they consumed peanuts).*"
        - "*No, it is not likely that factors other than bowl size could have explained the difference in the average amount of ice cream because we used random assignment, so all other variables are equalized because they have the same chance at being assigned to either group."*

Correct answers, but bringing in extraneous information (EXT)

- **C-EXT-RS Bringing up issues of generalization and/or random sampling extraneously when the question is about causation, while still correctly addressing the need for random assignment to make causal claims.** (e.g., when asked only about causal claims, says that we can make causal claims because the researchers used random assignment – but we cannot generalize to the population because the sampling was not random.)
    - Examples:
        - "*No, the researchers cannot generalize this statement because there was no random sampling that occurred, only random assignment."*
        - *"The [causal] claim is appropriate because random assignment allows causal claims to be made to the sample but not necessarily [to] the population of interest."*
- **C-EXT-RA Bringing up issues of causation and/or random assignment extraneously when the question is about generalization, while still correctly addressing the need for random sampling to make generalizations.** (E.g., when asked only about generalization, says "No we can't generalize because we don't have a random sample. Also, we cannot make causal claims because the assignment to groups was not random.)
    - Examples:
        - *[Being asked if a headline that makes a generalization can be published]: "The headline is appropriate to make a generalization because it was a random sample for adults 18 and older. It is not appropriate to make a causal claim because there was no isolated variables to conclude that drinking is the explanatory variable for*

improvement in emotional health. Therefore, random assignment was not included in the experiment."

- *[Being asked about a headline that makes a generalization]: "Yes, since random sampling was used we can make a generalization for this population. Since random assignment was not used in this study design, we can only make an associative claim which is indicated in the headline."]*

Ambiguity (Scorer may have difficulty judging whether or not student has a correct understanding) (A)

- **A-BOTH Does not separate generalization and causation, saying you need both random sampling and random assignment to conclude generalization and causation.** (E.g., saying that we cannot generalize or make causal claims because there was no random sampling nor random assignment.)
    - Example: "*Random sampling and random assignment is necessary to ensure the cause-and-effect relationship and to generalize to the entire population*."
- **A-RAND Being vague about what kind of randomness is needed to generalize or make causal claims** (E.g., just mentioning "random" but not being specific about random sampling or random assignment)
    - Example: "*The researchers can not [sic] make generalization or casual claims because no type [of] randomness was used*."
- **A-RSNORA Saying that only random sampling was used, thus implying that random assignment was not used** (e.g., saying that we cannot make causal claims because random sampling was used, without mentioning lack of random assignment.)
    - Example: "*In this example, they made a causal claim, but based on the study design which used only random sampling, only a generalization can be made, not a causal claim.*"
- **A-RANORS Saying that only random assignment was used, thus implying that random sampling was not used** (e.g., saying that we cannot generalize to the population because random assignment was used, without mentioning lack of random sampling.)
    - Example: "*Based on the study design (random assignment) the researchers cannot generalize their findings to this population because the sample, which consisted of only high-risk children, is not representative of <u>all</u> children*"

Appendix L1: Codes Specific to Lab Assignment

For question #13, which asks whether a hypothetical classmate is correct in saying that random sampling would allow a cause-and-effect conclusion, the following behaviors were coded:

- **I-LAB13-RSCC Incorrectly agreeing that random sampling allows for causal claims**
    - Examples:
        - "*Yes [the classmate is correct] because they used random sampling*."
        - "*Yes the classmate's statement is correct because the population was randomly sampled*."
- **C-LAB13-RSGEN Correctly mentioning that random sampling only helps with generalization**
    - Examples:
        - "*Random sampling would only allow us to generalize results of a study to a particular population*."
        - "*No his statement is not correct. A random sample would provide a generalization for the population, not causation.*"
- **C-LAB13-RACC Correctly mentioning that random assignment would be needed for making causal claims**
    - Examples:
        - "*No, the classmate is not correct because that would be a cause and effect statement and only random assignment, not random sampling, allows you to makes those kind of statements*."
        - "*I would tell my colleague that based on the study design we are not able to make causal claims because the mothers were not randomly assigned to eat peanuts or not eat peanuts*."

For question #14, which asks students whether or not they would advise a pregnant colleague to avoid eating peanuts based on results of a study that uses neither random sampling nor random assignment, the following behaviors were coded:

- **C-LAB14-NOCC Mentioning the lack of ability to make causal claims (or pointing out that random assignment was not used, or that confounding variables could explain peanut sensitivity)**
    - Examples:
        - "*Based on the design, I would tell her that there is only an association, and since they did not use random assignment in the study, we do not know that that is what caused the peanut allergy, since there could be many other confounding variables.*"
        - "*I would tell her that because there was no random assignment involved in the study, there is no way to infer that peanut consumption during pregnancy is what caused the infants to become sensitive to peanuts and that there could have been other factors involved*."

- **C-LAB14-NOGEN Mentioning the lack of ability to make generalizations (or pointing out that random sampling was not used, or that the sample may not be representative of the population)**
    - Examples:
        - "*Based on the design of the study I would tell her not to worry about it because the experiment didn't use random sampling the sample wasn't representative of the entire population and you can't make generalized statement*."
        - "*The study wasn't a random sample, and doesn't represent the population. Therefore that generalization can't be made*."
- **I-LAB14-PVAL Makes a decision based only on the *p*-value, without consideration of study design**
        - Example: "*Since the p-value of peanut allergy development and frequent consumption of peanuts during pregnancy was very low (meaning high support), I would support my coworker's decision to avoid eating peanuts during pregnancy*."
- **I-LAB14-NOSD Makes a decision based on factors not related to study design.**
        - Example:"*I would tell my colleague that she can go ahead and avoid peanuts if she so chooses because of the statistics presented in excerpt #2, but there is no guarantee that will improve her child's chances of not developing a peanut allergy because we do not have proof maternal consumption of peanuts during pregnancy causes the insensitivity.*"

Appendix L2: Codes specific to Group Quiz

For the scenarios in questions #1-2 and in questions #5-6, students were presented with potential newspaper headlines and asked about their appropriateness given the study design. As interpretation of the headlines was required to answer these questions, the following behaviors were coded:

- **I-QUIZ-HGEN Difficulty recognizing whether a headline is making a generalization**
    - Examples:
        - Q1: "*No, this study only shows correlation because there is no random assignment, so we cannot prove a causal relationship between these variables"* (Misinterpreting headline as causal claim, rather than as a generalization.)
        - Q6: "*Yes…this headline is more accurate because while higher grades are correlated with admission, they don't necessarily cause admission*." (Not mentioning anything about generalization made in the headline.)

- **I-QUIZ-HCC Difficulty recognizing whether a headline is making a causal claim**
    - Examples:
        - Q1: "*No. Because it's a random sample, you are able to generalize but you cannot make a causal claim because it's not random assignment."* (Misinterpreting a claim that generalizes an association as a causal claim.)
        - Q1: "*No, we cannot have causation because there was no random assignment, so we can't conclude that moderate drinking causes better emotional health.*" (Misinterpreting a claim that generalizes an association as a causal claim.)

Appendix M: Results from Qualitative Analysis Coding

Appendix M1: Lab Assignment coding

The lab assignment consisted of one single context: infants and peanut allergies, and described two studies. Therefore, the lab was examined and coded as a whole, although there were some codes that were specific to answers presented in the last two questions. The codes were developed using the following labels corresponding to categories and sub-categories:

I = incorrect thinking
- TC = Types of conclusions
- SS = Sample size
- SD = Study descriptions

C = correct thinking
- SG = Sampling and generalization
- AC = Assignment and conclusions
- WHY = Elaborating on why certain study designs lead to given conclusions
- EXT = Providing extraneous information

A = ambiguity (scorer may have difficulty judging whether or not student displays correct understanding)

(Note: In the tables below, these abbreviations are used: RS = random sampling, RA = random assignment.)

Code	Behavior	% of Section				% of all
		1 (n = 40)	2 (n = 31)	3 (n = 27)	4 (n = 30)	(n = 128)
[I]	*Misconceptions/Incorrect Thinking*					
[I-TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	15.0	22.5	33.3	30.0	24.2
I-TC-RSC	Bringing up only random sampling/lack thereof when the question is about causation	2.5	0.0	3.7	13.3	4.7
I-TC-RAG	Bringing up only random assignment/lack thereof when the question is about generalization	7.5	12.9	18.5	10.0	11.7
I-TC-BOTHG	Need both random sampling AND random assignment to generalize	2.5	3.2	11.1	3.3	4.7
I-TC-BOTHC	Need both random sampling AND random assignment to make causal claims	2.5	3.2	7.4	3.3	3.9

Code	Behavior	% of Section				% of all
		1 (n = 40)	2 (n = 31)	3 (n = 27)	4 (n = 30)	(n = 128)
I-TC-CLAIM	Confusing the meaning of "generalize" with the meaning of "causal claims"	2.5	3.2	0.0	10.0	3.9
I-TC-NOCC	Not believing causal claims can be made even though random assignment was used	0.0	3.2	0.0	3.3	1.6
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	2.5	0.0	0.0	13.3	3.9
I-SS-UNEVEN	Unequal sample sizes in two groups do not allow for any conclusions	0.0	0.0	0.0	0.0	0.0
I-SS-LARGEN	Large sample size allows for generalization	2.5	0.0	0.0	10.0	3.1
I-SS-SMALLN	Small sample size does not allow for any conclusions	0.0	0.0	0.0	3.3	0.8
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	15.0	12.9	14.8	20.0	15.6
I-SD-RECRS	Difficulty understanding whether RS was used	12.5	9.7	14.8	16.7	13.3
I-SD-RECRA	Difficulty understanding whether RA was used	2.5	6.5	0.0	3.3	3.1
[C]	*Correct Thinking*					
[C-SG]	*Makes connections between sampling and generalization: Either mentions lack of RS OR how sample is different from population*[a] *(at least one SG code)*	100.0	100.0	96.3	70.0	92.2
C-SG-RSGEN	Random sampling is relevant for generalization	95.0	93.6	88.9	66.7	86.7
C-SG-SCHAR	Mention that characteristics make sample different from population (if no RS used)	30.0	38.7	29.6	26.7	31.3
[C-AC]	*Makes connections between random assignment and causation. Either mentions lack of RA OR how groups are different from each other (confounding)*[b] *(at least one AC code)*	95.0	93.6	96.3	66.7	88.3
C-AC-RACC	Random assignment is relevant for causation	95.0	93.6	96.3	66.7	88.3
C-AC-CONFV	Mention that confounding variables can make groups	5.0	3.2	0.0	0.0	2.3

Code	Behavior	% of Section				% of all
		1 (*n* = 40)	2 (*n* = 31)	3 (*n* = 27)	4 (*n* = 30)	(*n* = 128)
	different from each other (if no RA used)					
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	57.5	51.6	48.2	23.3	46.1
C-WHY-RS	Explaining *why* random sampling allows for generalization	37.5	29.0	14.8	10.0	24.2
C-WHY-RA	Explaining *why* random assignment allows for causation	50.0	48.4	48.2	16.7	41.4
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	22.5	38.7	22.2	20.0	25.8
C-EXT-RS	Bringing up RS or generalization when question is about causation only - but still talking correctly about causation	15.0	35.5	11.1	13.3	18.8
C-EXT-RA	Bringing up RA or causation when question is about generalization only - but still talking correctly about generalization	10.0	9.7	14.8	16.7	12.5
[A]	*Ambiguity (at least one A code)*	5.0	9.7	14.8	20.7	11.8
A-BOTH	Saying you need RS and RA to generalize and make causal claims	2.5	9.7	11.1	6.7	7.0
A-RAND	Vagueness about "randomness" without specifying type of randomness.	2.5	3.2	3.7	13.3	5.5
A-RSNORA	"Cannot make causal claims because RS was used" only implying RA was not	0.0	0.0	0.0	0.0	0.0
A-RANORS	"Cannot make generalizations because RA was used" only implying RS was not	2.5	0.0	7.4	0.0	2.3
Question 13						
I-LAB13-RSCC	Says classmate is correct that RS leads to causation	5.0	6.5	0.0	23.3	8.6
[C-LAB13]	*Either explains RS is only for generalization, or explains need for RA for causation[c] (at least one C "correct" code for question #13)*	87.5	87.1	96.3	56.7	82.0
C-LAB13-RSGEN	Says RS is only for generalization	52.5	71.0	55.6	26.7	51.6

Code	Behavior	% of Section				% of all (n = 128)
		1 (n = 40)	2 (n = 31)	3 (n = 27)	4 (n = 30)	
C-LAB13-RACC	Correctly brings up need for RA for causation (or problems with confounding)	82.5	64.5	66.7	53.3	68.0
Question 14						
[C-LAB14]	*Either mentions lack of ability to make causal claims, or lack of ability to make generalizations (at least one C "correct" code for question 14)*	90.0	83.8	85.2	50.0	78.1
C-LAB14-NOCC	Mention lack of ability to make causal claims	80.0	83.9	85.2	36.7	72.9
C-LAB14-NOGEN	Mention lack of ability to generalize	52.5	32.3	51.9	33.3	42.3
I-LAB14-PVAL	Decision based only on p-value	2.5	6.5	3.7	6.7	4.7
I-LAB14-NOSD	Decision based on factors not related to study design or results	7.5	6.5	7.4	20.0	10.2

[a]The percentage of students who pointed out the lack of ability to make generalizations by either mentioning the lack of random sampling (C-SG-RSGEN) and/or mentioning that the sample is different in characteristics from the population was computed (C-SG-SCHAR). Either of these two approaches would constitute a correct approach.

[b]The percentage of students who pointed the need for random assignment to make causal claims (C-AC-RACC) and/or mentioning that confounding variables can explain differences between groups was computed (C-AC-CONFV). Either of these two approaches would constitute a correct approach.

[c]In question #13 on the lab, the percentage of students who pointed out that the classmate was incorrect because random sampling is for making generalizations (C-LAB13-RSGEN), and/or pointing out the need for random assignment to make causal claims was computed (C-LAB13-RACC). Either of these two approaches would constitute a correct answer.

Appendix M2: Coding of Group Quiz

The quiz consisted of three different scenarios. For each scenario, there was one question mainly related to generalization and one question mainly related to causation. Therefore, each set of two questions (i.e. each separate context) was coded for the group quiz. There were 43 total group quizzes coded.

Questions #1 and #2: Gallup poll on drinking and emotional health

Code	Behavior	% of groups per section				% of all groups (n = 43)
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	
[I]	*Misconceptions/Incorrect Thinking*					
[TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	14.3	8.3	0.0	25.0	11.6
I-TC-RSC	Bringing up only random sampling/lack thereof when the question is about causation	7.1	0.0	0.0	0.0	2.3
I-TC-RAG	Bringing up only random assignment/lack thereof when the question is about generalization	0.0	0.0	0.0	12.5	2.3
I-TC-BOTHG	Need both random sampling AND random assignment to generalize	0.0	0.0	0.0	0.0	0.0
I-TC-BOTHC	Need both random sampling AND random assignment to make causal claims	0.0	8.3	0.0	0.0	2.3
I-TC-CLAIM	Confusing the meaning of "generalize" with the meaning of "causal claims"	7.1	0.0	0.0	12.5	4.7
I-TC-NOCC	Not believing causal claims can be made even though random assignment was used	0.0	0.0	0.0	0.0	0.0
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	7.1	8.3	0.0	0.0	4.7
I-SS-UNEVEN	Unequal sample sizes in two groups do not allow for any conclusions	0.0	0.0	0.0	0.0	0.0
I-SS-LARGEN	Large sample size allows for generalization	7.1	8.3	0.0	0.0	4.7
I-SS-SMALLN	Small sample size does not allow for any conclusions	0.0	0.0	0.0	0.0	0.0
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	7.1	0.0	0.0	0.0	2.3

Code	Behavior	% of groups per section				% of all groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
I-SD-RECRS	Difficulty understanding whether RS was used	7.1	0.0	0.0	0.0	2.3
I-SD-RECRA	Difficulty understanding whether RA was used	0.0	0.0	0.0	0.0	0.0
[C]	*Correct Thinking*					
C-SG-RSGEN	Recognizes that random sampling is relevant for generalization (in this case, we have a random sample so we can generalize to a population) [a]	78.6	83.3	100.0	75.0	83.7
[C-AC]	*Makes connections between assignment and causation. Either mentions lack of RA OR how groups are different from each other (confounding)*[b] *(at least one AC code)*	92.9	83.3	77.8	100.0	88.4
C-AC-RACC	Random assignment is relevant for causation	78.6	58.3	77.8	87.5	74.4
C-AC-CONFV	Mention that confounding variables can make groups different from each other	64.3	33.3	33.3	62.5	48.8
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	21.4	0.0	0.0	37.5	14.0
C-WHY-RS	Explaining *why* random sampling allows for generalization	14.3	0.0	0.0	12.5	7.0
C-WHY-RA	Explaining *why* random assignment allows for causation	7.1	0.0	0.0	25.0	7.0
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	7.1	33.3	44.4	87.5	37.2
C-EXT-RS	Bringing up RS or generalization when question is about causation only - but still talking correctly about causation	0.0	0.0	11.1	62.5	14.0
C-EXT-RA	Bringing up RA or causation when question is about generalization only - but still talking correctly about generalization	7.1	33.3	33.3	50.0	27.9
[A]	*Ambiguity (at least one A code)*	0.0	8.3	22.2	0.0	7.0
A-BOTH	Saying you need RS and RA to generalize and make causal claims	0.0	0.0	0.0	0.0	0.0

Code	Behavior	% of groups per section				% of all groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
A-RAND	Vagueness about "randomness" without specifying type of randomness.	0.0	0.0	0.0	0.0	0.0
A-RSNORA	"Cannot make causal claims because RS was used" only implying RA was not	0.0	8.3	22.2	0.0	7.0
A-RANORS	"Cannot make generalizations because RA was used" only implying RS was not	0.0	0.0	0.0	0.0	0.0
Quiz-specific codes for items involving headlines						
I-QUIZ-HGEN	Not recognizing when headline is/is not making a generalization	21.4	8.3	0.0	12.5	11.6
I-QUIZ-HCC	Not recognizing when headline is/is not making a causal claim	14.3	16.7	22.2	0.0	14.0

[a]The study in question was designed with random sampling. Therefore, the code C-SG-SCHAR (mentioning characteristics that make sample different from the population) was not used, as it did not represent correct reasoning for this context.

[b] The study in question was designed without random assignment. The percentage of student groups who pointed out the lack of ability to make causal claims by either mentioning the lack of random assignment (C-AC-RACC) and/or mentioning that confounding variables can explain differences between groups (C-AC-CONFV) was computed. Either of these two approaches would constitute a correct approach.

Questions #3 and #4: Nutritionists and ice cream bowl sizes

Code	Behavior	% of groups per section				% of all Groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
[I]	*Misconceptions/Incorrect Thinking*					
[TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	35.7	25.0	0.0	37.5	25.6
I-TC-RSC	Bringing up only random sampling/lack thereof when the question is about causation	21.4	25.0	0.0	25.0	18.6
I-TC-RAG	Bringing up only random assignment/lack thereof when the question is about generalization	0.0	0.0	0.0	0.0	0.0
I-TC-BOTHG	Need both random sampling AND random assignment to generalize	7.1	0.0	0.0	0.0	2.3
I-TC-BOTHC	Need both random sampling AND random assignment to make causal claims	0.0	0.0	0.0	0.0	0.0

Code	Behavior	% of groups per section				% of all Groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
I-TC-CLAIM	Confusing the meaning of "generalize" with the meaning of "causal claims"	21.4	0.0	0.0	0.0	7.0
I-TC-NOCC	Not believing causal claims can be made even though random assignment was used	0.0	0.0	0.0	12.5	2.3
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	7.1	0.0	0.0	12.5	4.7
I-SS-UNEVEN	Unequal sample sizes in two groups do not allow for any conclusions	0.0	0.0	0.0	0.0	0.0
I-SS-LARGEN	Large sample size allows for generalization	0.0	0.0	0.0	0.0	0.0
I-SS-SMALLN	Small sample size does not allow for any conclusions	7.1	0.0	0.0	12.5	4.7
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	7.1	33.3	33.3	37.5	25.6
I-SD-RECRS	Difficulty understanding whether RS was used	7.1	8.3	11.1	0.0	7.0
I-SD-RECRA	Difficulty understanding whether RA was used	0.0	25.0	22.2	37.5	18.6
[C]	*Correct Thinking*					
[C-SG]	*Makes connections between sampling and generalization: Either mentions lack of RS OR how sample is different from population*[a] *(at least one SG code)*	92.9	83.3	88.9	87.5	88.4
C-SG-RSGEN	Random sampling is relevant for generalization	78.6	75.0	88.9	87.5	81.4
C-SG-SCHAR	Mention that characteristics make sample different from population	28.6	25.0	33.3	37.5	30.2
C-AC-RACC	Recognizes that random assignment is relevant for causation (in this case, we have random assignment so we can make causal claims) [b]	85.7	58.3	77.8	87.5	76.7
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	71.4	33.3	77.8	62.5	60.5

Code	Behavior	% of groups per section				% of all Groups (n = 43)
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	
C-WHY-RS	Explaining *why* random sampling allows for generalization	21.4	8.3	22.2	25.0	18.6
C-WHY-RA	Explaining *why* random assignment allows for causation	71.4	33.3	77.8	62.5	60.5
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	14.3	16.7	11.1	12.5	14.0
C-EXT-RS	Bringing up RS or generalization when question is about causation only - but still talking correctly about causation	0.0	0.0	0.0	0.0	0.0
C-EXT-RA	Bringing up RA or causation when question is about generalization only - but still talking correctly about generalization	14.3	16.7	11.1	12.5	14.0
[A]	*Ambiguity (at least one A code)*	0.0	8.3	0.0	12.5	4.7
A-BOTH	Saying you need RS and RA to generalize and make causal claims	0.0	0.0	0.0	12.5	2.3
A-RAND	Vagueness about "randomness" without specifying type of randomness.	0.0	8.3	0.0	0.0	2.3
A-RSNORA	"Cannot make causal claims because RS was used" only implying RA was not	0.0	0.0	0.0	0.0	0.0
A-RANORS	"Cannot make generalizations because RA was used" only implying RS was not	0.0	0.0	0.0	0.0	0.0

[a]The study in question was designed with random assignment, but not random sampling. The percentage of student groups who pointed out the lack of ability to make generalizations by either mentioning the lack of random sampling and/or mentioning that this sample may not accurately represent the population was computed. Either of these two approaches would constitute a correct approach.

[b]The code C-AC-CONFV used for other questions was not used for questions 3 and 4, because the study in question was an experiment with random assignment. Therefore, discussing that confounding variables make the groups different from each other would actually constitute an incorrect, not a correct, form of thinking.

Questions #5 and #6: GPA and Medical School Admissions

Code	Behavior	% of groups per section				% of all groups
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	(n = 43)
[I]	*Misconceptions/Incorrect Thinking*					
[TC]	*Misunderstandings about which study designs help with which types of conclusions (at least one TC code)*	0.0	0.0	11.1	12.5	4.7
I-TC-RSC	Bringing up only random sampling/lack thereof when the question is about causation	0.0	0.0	11.1	12.5	4.7
I-TC-RAG	Bringing up only random assignment/lack thereof when the question is about generalization	0.0	0.0	0.0	0.0	0.0
I-TC-BOTHG	Need both random sampling AND random assignment to generalize	0.0	0.0	0.0	0.0	0.0
I-TC-BOTHC	Need both random sampling AND random assignment to make causal claims	0.0	0.0	0.0	0.0	0.0
I-TC-CLAIM	Confusing the meaning of "generalize" with the meaning of "causal claims"	0.0	0.0	11.1	0.0	2.3
I-TC-NOCC	Not believing causal claims can be made even though random assignment was used	0.0	0.0	0.0	0.0	0.0
[I-SS]	*Incorrect beliefs about sample size (at least one SS code)*	0.0	0.0	0.0	12.5	2.3
I-SS-UNEVEN	Unequal sample sizes in two groups do not allow for any conclusions	0.0	0.0	0.0	0.0	0.0
I-SS-LARGEN	Large sample size allows for generalization	0.0	0.0	0.0	0.0	0.0
I-SS-SMALLN	Small sample size does not allow for any conclusions	0.0	0.0	0.0	12.5	2.3
[I-SD]	*Difficulty understanding study descriptions (at least one SD code)*	0.0	0.0	0.0	0.0	0.0
I-SD-RECRS	Difficulty understanding whether RS was used	0.0	0.0	0.0	0.0	0.0
I-SD-RECRA	Difficulty understanding whether RA was used	0.0	0.0	0.0	0.0	0.0
[C]	*Correct Thinking*					
C-SG-RSGEN	Recognizes that random sampling is relevant for generalization (in this case, we	78.6	66.7	77.8	50.0	69.8

Code	Behavior have a random sample so we can generalize to a population) [a]	% of groups per section				% of all groups (n = 43)
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	
[C-AC]	*Makes connections between assignment and causation. Either mentions lack of RA OR how groups are different from each other (confounding)*[b] *(at least one AC code)*	92.9	100.0	77.8	87.5	90.7
C-AC-RACC	Random assignment is relevant for causation	85.7	66.7	66.7	50.0	69.8
C-AC-CONFV	Mention that confounding variables can make groups different from each other	50.0	66.7	33.3	62.5	53.5
[C-WHY]	*Answer includes more depth: Student elaborates about why certain study designs lead to given conclusions (at least one WHY code)*	0.0	0.0	11.1	0.0	2.3
C-WHY-RS	Explaining *why* random sampling allows for generalization	0.0	0.0	0.0	0.0	0.0
C-WHY-RA	Explaining *why* random assignment allows for causation	0.0	0.0	11.1	0.0	2.3
[C-EXT]	*Correct answers, but bringing in extraneous information (at least one EXT code)*	42.9	91.7	33.3	25.0	51.2
C-EXT-RS	Bringing up RS or generalization when question is about causation only - but still talking correctly about causation	21.4	33.3	0.0	25.0	21.0
C-EXT-RA	Bringing up RA or causation when question is about generalization only - but still talking correctly about generalization	35.7	83.3	33.3	25.0	46.5
[A]	*Ambiguity (at least one A code)*	7.1	0.0	22.2	0.0	7.0
A-BOTH	Saying you need RS and RA to generalize and make causal claims	0.0	0.0	0.0	0.0	0.0
A-RAND	Vagueness about "randomness" without specifying type of randomness.	0.0	0.0	0.0	0.0	0.0
A-RSNORA	"Cannot make causal claims because RS was used" only implying RA was not	7.1	0.0	22.2	0.0	7.0
A-RANORS	"Cannot make generalizations because RA was used" only implying RS was not	0.0	0.0	0.0	0.0	0.0

Code	Behavior	% of groups per section				% of all groups (n = 43)
		1 (n = 14)	2 (n = 12)	3 (n = 9)	4 (n = 8)	
Quiz-specific codes for items involving headlines						
I-QUIZ-HGEN	Not recognizing when headline is/is not making a generalization	21.4	41.7	0.0	50.0	27.9
I-QUIZ-HCC	Not recognizing when headline is/is not making a causal claim	0.0	0.0	0.0	0.0	0.0

[a]The study in question was designed with random sampling. Therefore, the code C-SG-SCHAR (mentioning characteristics that make sample different from the population) was not used, as it did not represent correct reasoning for this context.

[b] The study in question was designed without random assignment. The percentage of student groups who pointed out the lack of ability to make causal claims by either mentioning the lack of random assignment (C-AC-RACC) and/or mentioning that confounding variables can explain differences between groups (C-AC-CONFV) was computed. Either of these two approaches would constitute a correct approach.